Unit 12 Review

Contents

Introduction

Another way of drawing material together

This unit is designed to help you consolidate what you have learned in M140. It describes some extensions of ideas you have already met, as well as a small number of new statistical ideas. The aim is to draw together parts of the module that are closely linked, even though they may have been in different units. Each section will review one or two topics and contain a number of activities to help you refresh your knowledge of them.

- Section 1 reviews descriptive statistics and summary statistics, which were the focus of Units 1, 2 and 3 (Book 1). There is new material in Subsection 1.2, which introduces *growth charts*.
- Section 2 discusses the collection of data through surveys and in experiments, including clinical trials. The material is based mainly on Units 4, 10 and 11, with the addition of material on combining survey methods (Subsection 2.1).
- Section 3 reviews properties of probability given in Units 6 and 8. Also, in Subsection 3.2, the general form of the binomial distribution is introduced. (Unit 6 used a specialised form of this distribution.)
- Section 4 describes the principal steps in a hypothesis test (Unit 6) and considers the χ^2 test of independence in a contingency table (Unit 8).
- Section 5 reviews the properties of the normal distribution and examines hypothesis tests and confidence intervals for making inferences about the mean of a population or the difference between two population means (Units 7, 9 and 10). A two-sample t-test for populations with unequal variances is introduced.
- Section 6 concerns relationships between two variables and reviews regression and correlation, which were the main focus of Units 5 and 9.
- Section 7 uses Minitab to explore binomial distributions and perform two-sample t-tests.

In planning your study, you should note that there is new (assessable) material in Subsection 1.2, a small part of Subsection 2.1, most of Subsection 3.2 and a large part of Subsection 5.2.

Section 7 directs you to the Computer Book. You are also guided to the Computer Book at the end of Sections 3 and 5. It is better to do the work at those points in the text, although you can leave it until later if you prefer.

1 Summarising data

One reason for summarising data is to be able to report the data succinctly, perhaps quoting its median value or range in order to describe features of the data. As well as numeric summaries, figures such as a boxplot or stemplot can be used for this purpose and are very informative. Another important reason for summarising data, which you saw in later units, is that summary statistics are often all that are needed for performing hypothesis tests or calculating confidence intervals. In this context there is little choice as to how the data should be summarised. For example, the sample mean and the sample size are the information required from the data in order to perform a one-sample z-test.

In Subsections 1.1 and 1.2, numerical statistics for summarising data are described. In Subsection 1.3, we turn to graphical summaries. In Subsection 1.4, we focus on the Retail Prices Index (RPI) and other indexes for summarising data on prices and earnings. All these topics are primarily taught in Units 1 to 3.

1.1 Common numerical summaries

Numbers that are used to summarise data are referred to as *summary statistics*. Usually, the key quantities used to summarise a set of data are its median or mean, along with its interquartile range, standard deviation or variance.

Median and mean

The median is the middle item in a set of data (if the number of items in the batch is odd) or the average of the middle two items (if there are an even number of items in the batch).

The mean is the average value in a set of data, given by:

$$\overline{x} = \frac{\sum x}{n},$$

where n is the number of items in the set of data.

If you have to give just *one* number to summarise a set of data, then the median or mean are the obvious choices – one gives the middle of the data and the other its average. The two are quite close if the data have a fairly symmetric distribution, but will differ more if the data show great skewness. When the data are highly skew, the median is often more representative of the data.

0	5
1	2
2	9
3	1 3 7
4	0 1 4 6
5	2 6 7 9
6	4 8 9
7	1 7
8	6
9	3 5

$n = 22$ 0 | 5 represents 5

Reasonably symmetric

4	4 4 5 5 5 5 7 7 8
5	1 2 3 6 8 9
6	0
7	4 8
8	2
9	
10	
11	4

$n = 20$ 4 | 4 represents 0.44

Highly skew

Figure 1 Examples of reasonably symmetric and highly skew datasets

Activity 1 Time between elections

Following the *Fixed-term Parliaments Act 2011*, general elections in the UK take place every five years. Before that, an election could be called at any point the Prime Minister wished.

The following are the times (in months) between elections in the years 1970 to 2010:

$$44, \quad 7, \quad 55, \quad 49, \quad 48, \quad 58, \quad 61, \quad 49, \quad 47, \quad 60.$$

(a) What is the median of these data?

(b) What is the mean of these data?

(c) Which observation is most responsible for the difference between the median and mean? Would you consider either the mean or median unrepresentative of these data?

The median and mean each describe the *location* of a set of data. They give a value that, in some sense, the data are centred around. The interquartile range, standard deviation and variance are each a measure of the extent to which a set of data is spread out.

Interquartile range

The central half of a set of data lies between the lower quartile and the upper quartile. The distance between these quartiles is the interquartile range.

In the ordered set of data, the lower quartile is at position $(n+1)/4$ and the upper quartile is at position $3(n+1)/4$.

Activity 2 Interquartile range examples

The following two sets of data have each been ordered from lowest to highest. The first set contains 15 data values and the second contains 20 data values. Determine the interquartile range of each set.

(a) 47, 49, 56, 57, 58, 58, 63, 63, 63, 64, 64, 65, 66, 68, 73.

(b) 13, 15, 16, 16, 17, 19, 20, 20, 20, 20, 21, 21, 21, 22, 23, 24, 26, 27, 28, 29.

Variance and standard deviation

The variance is the squared differences between each data value and the sample mean, added together and divided by $n-1$ (where n is the sample size).

The standard deviation is the square root of the variance.

Although, from its definition, the formula for the sample variance (s^2) is

$$s^2 = \frac{\sum(x-\overline{x})^2}{n-1},$$

it is quicker to separately calculate $\sum x^2$ and $\sum x$ and apply the equivalent formula

$$s^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right).$$

Activity 3 Lowestoft daily temperatures

The following are the mean daily maximum temperatures (in °C) in Lowestoft for July, in the years from 2002 to 2010:

$$20.5,\ 21.6,\ 19.6,\ 19.9,\ 23.7,\ 20.8,\ 21.1,\ 21.2,\ 23.6.$$

Determine the variance and the standard deviation of these data.

Victoria Beach, Lowestoft

A disadvantage of the variance is that its scale is not the scale of the original data. For example, suppose the data are the times taken to perform a task and each of these times is typically within 15 seconds of 2 minutes. Then the standard deviation might be, say, 10 seconds. This quantity is readily interpreted and can be related to the data values. In contrast, the variance would be 100 seconds^2, which cannot be related to the data values quite so easily.

The reason the variance is important is that it has good mathematical properties. For instance, to perform a two-sample z-test (Unit 7) requires ESE: the estimated standard error of $\overline{x}_A - \overline{x}_B$. This is calculated, not from the two standard deviations s_A and s_B, but from the variances s_A^2 and s_B^2:

$$\text{ESE} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}.$$

When a set of data is to be summarised by giving one quantity to indicate its location and another to indicate its spread, we generally give either the mean and standard deviation, or the median and interquartile range. Quoting other pairings, such as the mean and interquartile range, is less common.

The quantities used most commonly in statistical calculations (such as when forming confidence intervals or testing hypotheses) are means, variances and standard deviations. Even when testing whether a population median takes a specified value, we do not need the sample median.

1.2 Numerical summaries used less frequently

This subsection contains new material, not previously covered in M140.

The largest and smallest values in a dataset, E_U and E_L, are quite often recorded in conjunction with other summary statistics so as to give a fuller description of the data. (For example, they are included in five-figure summary tables.) Inspecting the values of E_U or E_L is a useful step in cleaning data prior to statistical analysis as unusual values can highlight major errors in a dataset, perhaps caused by typing errors or other errors in recording the data.

In contrast, the range, $E_U - E_L$, is seldom of interest because it is heavily influenced by the odd high or low value, so is often a poor reflection of the typical spread in a dataset. Similarly, the **mid-range**, $(E_u + E_L)/2$, and mode are seldom used as summary statistics even though they could be used as measures of location – they are generally less representative of the centre of the data than the mean or median, so one of the latter is used instead.

An informative way of giving a detailed summary of a large set of data is to identify some of its percentiles. As well as the median (50th percentile), lower quartile (25th percentile) and upper quartile (75th percentile), some of the deciles (10th, 20th, ..., 90th percentiles) might also be given. Also, in forming

confidence intervals and prediction intervals, the $2\frac{1}{2}$ and $97\frac{1}{2}$ percentiles are important as they are the end-points of a 95% interval. While confidence intervals summarise the results of a statistical analysis, rather than simply summarising a set of data, they illustrate that the more extreme percentiles can be of interest. Hence, it is often useful to include percentiles of less than 10% and more than 90% in a detailed summary of data.

To avoid a reader being swamped with numbers, the information from a large number of percentiles might be presented in a diagram. Figure 2 gives an example called a **growth chart**. It shows the distribution of weights in a very large sample of boys in the first year of life. The percentiles that are given are the 0.4th, 2nd, 9th, 25th, 50th, 75th, 91st, 98th and 99.6th.

The type of chart in Figure 2 is given to mothers leaving hospital after giving birth. It should reassure the majority of new parents that their baby is growing normally, while hopefully ringing alarm bells when a baby's weight is unusually high or low. Reading values from the graph shows, for instance, that at 28 weeks old:

- 25% of boys are below 7.5 kg (and 75% are above that weight).
- 9% of boys are below 7 kg.
- 2% are above 10 kg.

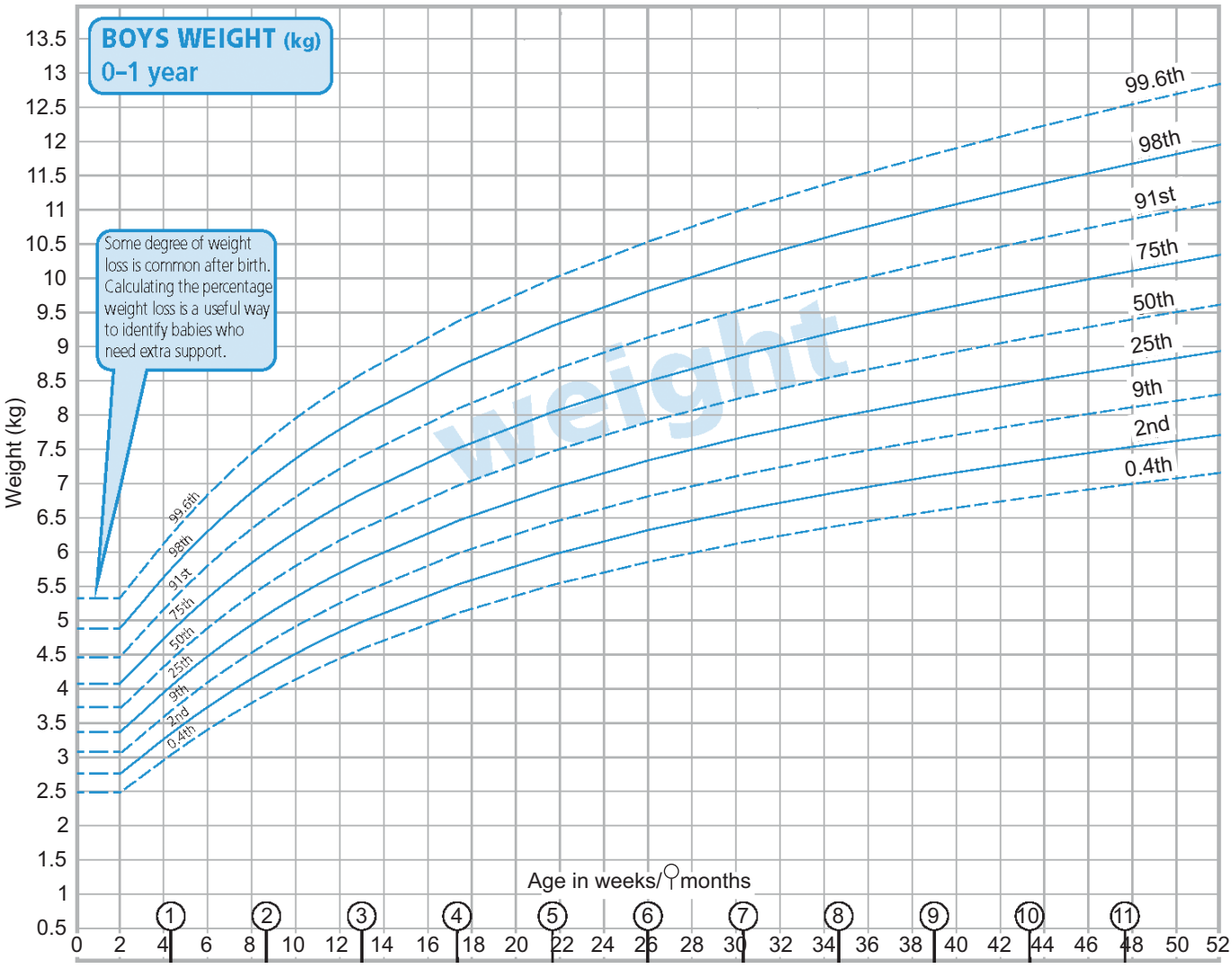


Figure 2 Growth chart showing percentiles of boys' weights in first year of life

Activity 4 Percentiles from a growth chart

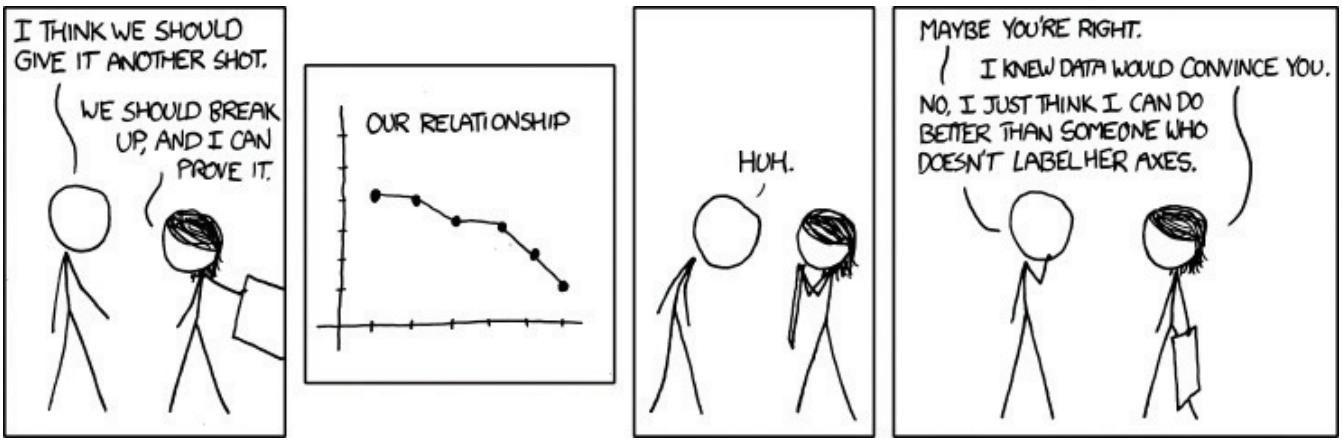
Use Figure 2 to give the proportion of boys who weigh:

(a) less than 4.5 kg when 10 weeks old;

(b) more than 10 kg when 46 weeks old;

(c) between 6 kg and 7.5 kg when 22 weeks old.

You have now covered the material related to Screencast 1 for Unit 12 (see the M140 website).



1.3 Graphical summaries

Boxplots and stemplots are commonly used as graphical summaries of data. If there are no unusually large or small data, a boxplot gives precisely the same information as a five-figure summary, except that the boxplot does not give n, the sample size.

Five-figure summary

	M	
n	Q_1	Q_3
	E_L	E_U

n batch size
M median
Q_1 lower quartile
Q_3 upper quartile
E_L lower extreme
E_U upper extreme

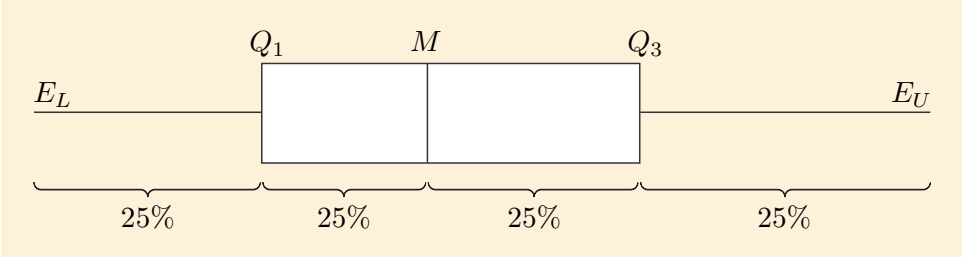


Figure 3 A standard boxplot

Activity 5 Five-figure summary and boxplot

The following data were given in part (a) of Activity 2 (Subsection 1.1):

$$47,\ 49,\ 56,\ 57,\ 58,\ 58,\ 63,\ 63,\ 63,\ 64,\ 64,\ 65,\ 66,\ 68,\ 73.$$

(a) Produce a five-figure summary of the data.

(b) Produce a boxplot of the data.

(c) Explain whether the boxplot indicates marked skewness in the data.

Details for drawing boxplots are given in Subsection 2.2 of Unit 3.

If there are very large or very small values (relative to the main body of data), then in a boxplot these are marked individually and the whiskers only extend as far as the lower adjacent value (for the left whisker) or the upper adjacent value (for the right whisker).

A stemplot resembles a histogram that has been turned on its side. It shows the shape of the distribution and contains most (or all) of the information in the data. Examples of stemplots are given in Figure 1 (Subsection 1.1). The following is a stemplot of the data in Activity 5.

4	7 9
5	
5	6 7 8 8
6	3 3 3 4 4
6	5 6 8
7	3

$n = 15$ 4 | 7 represents 47

From this stemplot we could give each value with full accuracy – no information is lost. In Activity 6, a little information is lost because the initial data are given to two decimal places, while the stemplot only gives one decimal place.

Activity 6 Olympic times for the 800 metres

The following data give the times of 24 athletes in the semi-finals of the men's 800-metre race at the 2012 Olympic Games. (A 25th runner was disqualified.) The times are given in seconds above 1 minute. For example, the fastest time of 1 minute 44.34 seconds is given below as 44.34 seconds. The data values have been ordered from fastest to slowest.

Table 1 800-metre race times, in seconds above 1 minute

44.34	44.35	44.51	44.54	44.63	44.74	44.87	44.93
45.08	45.09	45.10	45.34	45.44	45.63	45.84	45.85
46.14	46.19	46.29	46.66	47.52	48.18	48.98	53.46

(Data source: Official website of the Olympic Movement)

(a) Construct a stretched stemplot of the data, in which each whole second is split between two levels, and one outlier is listed separately.

(b) Comment on the shape of the stemplot.

In summarising data, the following is the order of priority.

1. If data must be summarised by just one number, then a number that represents the location of the data should be given (usually the median or mean).
2. If two numbers are to be used as the summary, then the second number should indicate the spread of the data (usually the interquartile range, standard deviation or variance).
3. Additional information would describe the shape of the data, notably any skewness, and identify the largest and smallest data along with any numbers that are extreme relative to the main body of the data.

Graphical summaries can convey a lot of information in an accessible way.

1.4 Summarising changes in prices and earnings

The Retail Prices Index (RPI) and Consumer Prices Index (CPI) are both used to summarise the overall change in the level of prices paid by people for the goods and services they buy. These price indexes are calculated in similar ways, and here we will focus on the RPI. Both use a very large 'basket of goods' that is designed to reflect the pattern of spending in the UK. Figure 4 shows the make-up of the RPI basket in 2012.

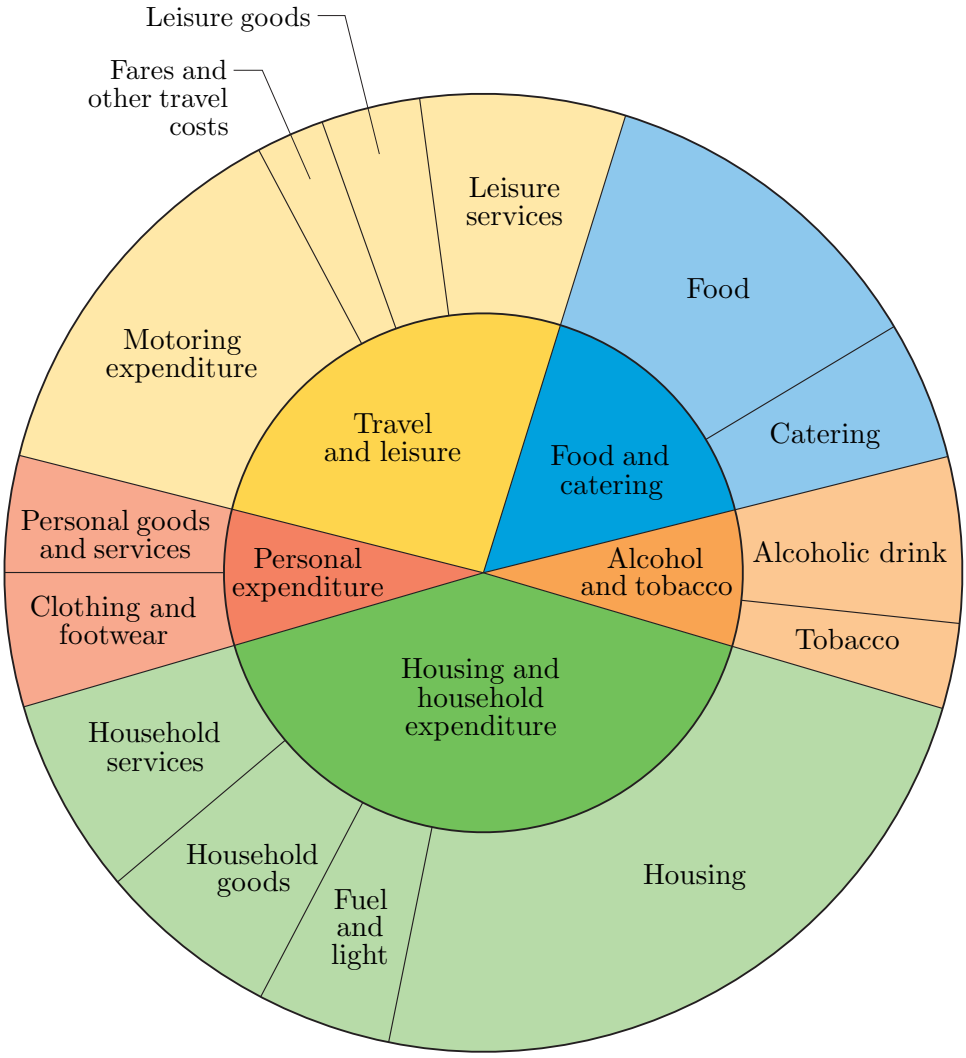


Figure 4 Structure of the RPI in 2012 (based on data from the Office for National Statistics)

As can be seen, the RPI is divided into five broad groupings. The inner ring shows, for example, that the typical household spends about twice as much on the group 'Food and catering' as on 'Personal expenditure'. The five groupings are divided into 14 more detailed subgroups, which are themselves divided into sections.

Certain items within each section are priced. For instance, within the 'Food and catering' group there is a 'Bread' section and the prices of representative items of bread (such as a large white sliced loaf and bread rolls) are monitored each month in a number of shops and supermarkets. For each item, its prices in the current month are compared with its prices in the previous January and a *price ratio* is calculated that fairly reflects how the price of the item has changed across the country.

Weighted mean of price ratios

The weighted mean of two or more numbers is:

$$\frac{\text{sum of } \{\text{number} \times \text{weight}\}}{\text{sum of weights}}.$$

So the weighted mean of two or more price ratios is:

$$\frac{\text{sum of } \{\text{price ratio} \times \text{weight}\}}{\text{sum of weights}}.$$

1. First, the price ratios of items within a subgroup are combined by taking their weighted mean – using weights that reflect expenditure patterns for the different items. These give a price ratio for each subgroup.
2. Next, the price ratios of subgroups within a group are combined by taking their weighted mean, giving a price ratio for the group.
3. Lastly, group price ratios are combined by taking their weighted mean, giving the *all-item price ratio* for that month.

The weights are determined from survey data on people's expenditures. They are set each January and used for a year. Calculating the all-item price ratio from the group price ratios is illustrated in Activity 7.

Activity 7 All-item price ratio for 2013

Group price ratios (r) for August 2013 relative to January 2013 are given in Table 2, where the weights (w) for 2013 are also given. Complete the last column of the table and show that the sum is 1021.086. Hence show that the all-item price ratio for August 2013 (relative to January 2013) is approximately 1.021.

Table 2 Weights and price ratios for August 2013

Group	Price ratio for August 2013 relative to January 2013 (r)	2013 weights (w)	Price ratio $\times$ weight ($r \times w$)
Food and catering	1.013	163	
Alcohol and tobacco	1.021	91	
Housing and household expenditure	1.017	419	
Personal expenditure	1.055	83	
Travel and leisure	1.022	244	

(Source: Office for National Statistics)

The RPI in January 2013 was derived from the RPI in January 2012, which in turn was derived from the RPI in 2011, and so on. Hence the RPI is a chained index. To give the chain a starting point, the RPI is set equal to 100 at a base date. The current base date for the RPI is January 1987.

The RPI in any month is obtained through multiplying 'that month's price ratio relative to the previous January' by 'the RPI in the previous January'. For example:

- The price ratio for January 2013 relative to January 2012 (the previous January) was 1.0328 and the RPI in January 2012 was 238.0. Hence the RPI in January 2013 was $1.0328 \times 238.0 \simeq 245.8$.
- As the price ratio for August 2013 relative to January 2013 was 1.021 (from Activity 7), the RPI in August 2013 was $1.021 \times 245.8 \simeq 251.0$.

The change in a person's annual expenditure seldom reflects change in prices, which is the change that the RPI aims to capture. Rather, the change in a person's annual expenditure probably reflects a change in the person's annual income or a change in their personal circumstances. It is for this reason that the RPI requires a notional basket of goods. In contrast, finding the change in a person's annual earnings simply requires knowledge of their earnings. Moreover, information on a person's earnings is usually carefully recorded – a by-product of the UK system of income tax. Hence, constructing an index of earnings does not face the same challenges that the RPI must overcome. There are, however, adjustments that surveys of earnings must make. For example, the Average

Weekly Earnings (AWE) index seasonally adjusts figures to allow for the effect of changes in earnings that occur regularly at fixed times of the year. Another example is the Annual Survey of Hours and Earnings (ASHE), which only collects information on paid employees and must make adjustments for the self-employed and unemployed, amongst others. Some detail is given in Unit 3.

2 Collecting data

Throughout this module we have examined samples of data. Sometimes the purpose of a sample is to learn about the population from which it comes, so the sample needs to be representative of the whole population and will usually need to be diverse. Another reason for gathering sample data is to perform an experiment. Then, often, the items selected for the experiment should be as similar as possible, so that differences between treatments (say) are not obscured by random variation between items.

Simple random sampling is the most common method of sampling, and many methods of testing hypotheses or forming confidence intervals assume that the data are a simple random sample from the population of interest. This is true of the *sign test* in Unit 6 where in one activity (Activity 29, Subsection 4.1) we have a random sample of 15 large schools from the East of England region, in another (Activity 34, Subsection 5.1) we have a random sample of secondary school academies from the East of England, and so forth. Similarly, random samples are used in one-sample z-tests (Unit 7, Activity 26 in Subsection 5.2 and Exercise 17 in Section 5, for example) and to form confidence intervals from z-tests (Unit 9, Activity 18 in Subsection 4.2). In order to compare the means of two populations it is common to take a simple random sample from each population, as illustrated in Unit 7 (Exercises 20 and 21, Section 6) and Unit 10 (Exercise 4, Section 3).

In principle, a simple random sample should be picked by sampling at random from the target population. This requires a list of the target population (or a mechanism that can select members of the target population at random) and is often impractical or impossible. Often a sample is treated as a simple random sample because it was not selected in any special way. For example, in your experiment with mustard seeds in Unit 10, the seeds are treated as a random sample of mustard seeds even though you did not use random number tables to decide which shop to go to for the seeds, or which packet of seeds to buy when you were in the shop.

In Subsection 2.1, we review survey methods that aim to find out about a population by examining just some of the items in the population, or questioning just some of the people in it. In Subsection 2.2, we discuss the collection of data for experiments.

2.1 Survey methods

When a large sample is to be gathered and no list of the population is available, then quota sampling is commonly employed so that characteristics of the population, such as age and gender, are reflected in the sample. For example, interviewers may be told to complete questionnaires with, say, thirty men aged 30–35, forty women aged 50–60, and so forth. This method might even be used when a list of the population is available, as it can be an economical method of gathering reasonably representative data. However, when a list of population members is available, other efficient survey methods can be

employed. Also, given a suitable list, proper randomisation can be used to give a simple random sample. The following methods were discussed in Unit 4.

Simple random sampling

In simple random sampling, members are selected one at a time. At each selection, those members of the population who have not already been selected are each equally likely to be the one selected. Thus each selection is independent of earlier selections, except that no member of the population can be selected more than once. The selection might be based on numbers given by a random number table or, more commonly, by a computer's randomisation procedure. In Unit 6, the module team used random numbers generated by Minitab to select a sample of schools from a list of schools in the East of England.

Systematic random sampling

Choosing a sample from a list of the target population is slightly easier using systematic random sampling rather than simple random sampling. If, say, one-seventh of the population is to be included in a systematic sample, then every seventh person in the list would be included in the sample, starting from one of the first seven people in the list, chosen at random. The random choice of starting point means that everyone in the population has an equal chance of being included in the sample.

If the population is listed in an order such that similar items or people are grouped together, then systematic sampling may well produce a more representative cross-section of the population than would be obtained by simple random sampling. For example, in an alphabetical listing of people, a husband and wife may well be listed such that one is immediately after the other. Then they would not both be included in a systematic sample, but through chance they might both be included in a simple random sample – which would then over-represent their family.

The following activity is designed to refresh your skills in using random number tables to choose samples for the above two types of survey.

Activity 8 Random and systematic samples

Table 3 gives the initials of 68 people who form the target population. The people are grouped in six sets (A, B, C, D, E and F) and have been labelled $01, 02, \ldots, 68$.

Table 3 A population divided into sets A–F

Label	Initials	Set	Label	Initials	Set	Label	Initials	Set
01	C.J.	A	24	R.D.M.	C	47	Z.G.	D
02	R.A.B.	A	25	M.C.	C	48	Y.H.	D
03	A.P.D.	A	26	E.L.	C	49	K.V.M.	E
04	M.A.	A	27	T.P.H.	C	50	P.H.	E
05	E.M.	A	28	I.M.	C	51	J.P.R.	E
06	J.L.	A	29	J.S.	C	52	G.C.T.	E
07	J.D.H.	A	30	C.G.	C	53	D.S.P.	E
08	A.E.G.	A	31	T.R.F.	C	54	H.M.	E
09	E.M.H.	B	32	C.C.T.	C	55	M.J.P.	E
10	T.N.	B	33	D.J.S.	C	56	C.S.T.	F
11	C.T.L.	B	34	S.G.C.	C	57	A.Y.K.	F
12	B.W.S.	B	35	D.L.	C	58	A.H.S.	F
13	J.A.R.	B	36	D.K.B.	C	59	D.B.M.	F
14	J.S.R.	B	37	C.A.M.	C	60	M.A.T.	F
15	S.L.	B	38	R.I.J.	C	61	A.N.D.	F
16	W.N.R.	B	39	W.O.J.	C	62	J.R.H.	F
17	A.C.D.	B	40	A.H.D.	D	63	R.J.C.	F
18	P.J.G.	B	41	P.V.	D	64	G.T.W.	F
19	W.W.S.	B	42	A.L.	D	65	G.K.S.	F
20	H.T.	B	43	L.R.P.	D	66	E.D.	F
21	G.B.Y.	B	44	D.A.F.	D	67	Y.S.H.	F
22	D.W.	B	45	R.H.R.	D	68	M.B.	F
23	M.S.	B	46	A.T.	D			

(a) Calculate the proportion of the population in each set.

(b) Choose a simple random sample of size 17 from the population in Table 3 using the random number table in the appendix to Unit 4, starting at the beginning of row **30**.

How many people in the sample are from set A? How many from each of the other sets? In the sample, which sets are under-represented relative to their size?

(c) Choose a systematic random sample of about a quarter of the population in Table 3. This time, take the first digit in row **10** in the range 1 to 4 as your random start. Analyse the sample with respect to 'set' and comment on how representative it is of the population.

(d) Explain whether you expected the sample obtained in (c) to represent the population better than the sample obtained in (b)?

Stratified sampling

Sometimes a population divides naturally into separate categories/ subpopulations, and items in the same category are likely to be more similar with respect to a quantity of interest than items from different categories. Then the categories are *strata*, provided each member of the population falls in exactly one category and we know (before gathering sample data) which members are in each category.

The approach in stratified sampling is to take a subsample from each stratum and combine the information the subsamples yield. If the quantity of interest is a numerical measurement on an interval scale (such as a length or weight, say), then the information from subsamples is combined as follows to yield an overall estimate of the population mean.

1. Taking each stratum separately, the sample mean and variance of the data in its subsample are calculated. These are estimates of the mean and variance for that complete stratum.
2. A weighted average of the strata means is calculated to obtain an estimate of the population mean. The weights are based on the number of data in each strata. (Formulas for the weights and the standard error of the weighted average are outside the scope of M140.)

If the quantity of interest is a proportion (the proportion of the population who prefer brand X to brand Y, say), then the proportion in each subsample is determined first – to give an estimate of the proportion in each stratum. A weighted average of these proportions is then calculated and taken as the overall estimate of the population proportion.

Cluster sampling

Cluster sampling is almost essential when data are to be gathered through face-to-face interviews and the population of interest is spread across a large geographical area or consists of a large number of locations. Many surveys require interviewers to call at people's homes or workplaces and cluster sampling is a way of reducing the travel involved in conducting the survey. In cluster sampling:

- The large geographical area is divided into small geographical areas – the *clusters*. (Each cluster might be an office block, or a street, say.)
- A limited number of these clusters are selected, preferably at random from all the clusters.
- Cluster samples are obtained by taking a random sample from each selected cluster. (For example, 20% of the people in each cluster might be questioned in the survey.)
- The cluster samples are combined to form a sample from the population.

In forming cluster samples, it is quite common to choose sample sizes so that the same proportion of each cluster is sampled. For example, the survey design might specify that 20% of the people in each selected cluster should be questioned in the survey. As long as the clusters that will be sampled are selected at random, each member of the target population will then have an equal chance of being included in the survey – which seems an attractive property. However, if the clusters in the population differ radically in size, a drawback of this approach is that the total sample size (and hence the cost of the sample) is not known until after the clusters to sample have been selected. There are variants of cluster sampling that are designed to avoid this problem.

Cluster sampling?

With stratified sampling, the strata subsamples must contain the same proportion of each stratum if each member of the target population is to have an equal chance of being sampled. Again, this is quite commonly done in practice, but there can be reasons for preferring alternatives. In particular, from previous surveys it may be known that certain strata display far greater variability than other strata. That is, the variance of the quantity of interest is known to be greater in particular strata. Those strata should then be sampled more heavily than strata in which the variability is less.

Combining survey methods

This is new material, not previously covered in M140.

In practice, it is common to combine different survey methods in designing a survey. To give an example, suppose a sample of staff working in hospitals across the UK is required. If it is thought that regional variations are likely, then the following survey design might be appropriate.

- Divide the UK into regions: say Scotland, Northern England, Wales, the Midlands and so on. Each region is a stratum and hospital staff from each stratum must be included in the sample.
- Within each region there are a lot of hospitals and so, to reduce the travelling a survey interviewer must do, each hospital might be treated as a cluster. Then, within each stratum, cluster sampling would be used to determine which hospitals the interviewer would visit.
- Suppose the survey must question nurses, doctors and administrators.

  To ensure a balanced sample is taken, these three categories of staff might be treated as three strata, and a random sample taken from each.

The above survey design might seem unnecessarily complicated, but surveys can be expensive. Moreover, large surveys are often repeated at regular intervals, so efficient and effective design is important. In Unit 2, some information was given about the survey methods used to obtain data for the Retail Price Index (RPI). For the RPI, the UK is divided into twelve regions (strata) and shopping locations in each region are placed within size categories.

- The size categories (another level of strata) are based on factors such as the size of the shopping population, the number of shops, and the drive-time to the shopping centre.
- Location selection takes place separately within each region, using a form of systematic sampling within strata. The outlets in a selected location are listed and the commodities each outlet sells are coded. The outlets are sampled in such a way as to obtain a number of prices for each item in the large basket of goods on which the RPI is based. (Some items in the basket of goods are sampled centrally, rather than through this procedure, but only for about 140 items out of 700.)

This procedure sounds complicated . . . and the detail around the separate sampling stages is substantially more complicated!

Activity 9 Sampling an urban area for interview

The Social Services department of a large unitary authority wishes to investigate whether disabled people in its urban area are receiving sufficient support. They plan to carry out a sample survey and decide to use as a sampling frame a list prepared for the purpose of collecting the council tax. This list includes every house in the urban area with its address, the house's band for council tax purposes (which is a measure of its estimated selling price), whether the house is owned privately or by the council, and the parish (the urban area is divided into 27 parishes) in which it is situated. The list contains about 60 000 houses.

The Social Services department wants to select a sample of about 3000 houses and will send an interviewer to each selected house. They will ask if any disabled people live at the house and, if so, ask about the support they receive.

(a) Giving reasons for your choices, describe how the Social Services department might use a procedure involving some forms of stratified

sampling, cluster sampling and simple random sampling to select the sample of 3000 houses. (It is not necessary to use all the methods.)

(b) State whether your procedure would be likely to increase or decrease the sampling error compared with that of a simple random sample of 3000 houses. What is the main benefit of your proposal compared with simple random sampling?

2.2 Collecting data in experiments

An experiment involves making specific observations under specific conditions in order to answer specific questions. There are many kinds of experiments, including:

- exploratory (Baconian) experiments, which aim to answer questions such as 'What happens if ... ?'
- measurement experiments
- hypothesis-testing experiments, which aim to test a specific hypothesis – often one about the cause of a phenomenon.

A Bacon(ian) experiment

In M140 we have concentrated on the third type of experiment. Most of the experiments have investigated the effect of a particular treatment of some sort on people, animals, plants or some kind of object. They include whether the roots of mustard seeds grow more in the light than in the dark; whether the weight gain on diet A differs from that on diet B; whether a new drug is more effective than a placebo; and which dose of drug gives the best combination of effectiveness and low risk of adverse side effects. The items or individuals from a population that are included in an experiment are referred to as the *experimental units*. The investigations in M140 often involved comparing experimental units that had been exposed to the experimental treatment (the experimental group) with other experimental units that had not been exposed to the treatment (the control

group). Otherwise, the experiments typically involved experimental groups that had been exposed to one of two treatments.

In designing experiments, one aim is to give fair comparison of the different treatments being examined.

You have met various strategies that help achieve this aim.

- Randomisation is the most important of these strategies. This allocates treatments to experimental units by chance, so no treatment can be deliberately favoured by being tested on the more responsive units. Hence, for example, in the mustard seed experiment you tossed a coin to decide which of the two pots of seeds would grow in the dark (step 12 in Subsection 2.3 of Unit 10).
- Apart from the characteristic being examined, the different treatments are made to resemble each other as much as possible. Thus placebos are used in clinical trials (and could be used in other forms of trial) so that 'treatment' and 'no treatment' appear the same to patients. For a similar reason, in the mustard seed experiment, the group of seeds grown in the light were covered with a piece of clear plastic so that they experienced similar levels of humidity as the seeds grown under aluminium foil.
- *Double-blind trials* are used so that knowledge of which treatment a patient is receiving cannot influence a patient's response or a clinician's perception of that response. This again aids a fair comparison of treatments.

Activity 10 Double-blind trials and placebos

Very briefly explain which of the following statements about double-blind trials are true and which are false.

A. In a double-blind clinical trial, as many as possible of those carrying out the trial, and of the patients receiving the treatment, must know which patients have received the active drug and which the placebo.

B. In a clinical trial where all measurements being made are objective rather than subjective, it is not necessary to use double-blind procedures.

C. In a clinical trial where all measurements being made are subjective rather than objective, it is not necessary to use double-blind procedures.

D. In a double-blind trial, the doctors who are administering the placebo and the patients who are receiving it all know that it is the placebo, whereas the doctors who are administering the drug begin tested and the patients who are receiving it do not know whether they are dealing with the drug or the placebo.

E. As far as possible, all controlled clinical trials should be double-blind.

F. In a double-blind trial, the patients should never be told of the possibility that they might receive a placebo.

G. In a double-blind clinical trial, neither the patients nor the doctors know which patients have received the drug being tested and which the placebo. However, an appropriate independent person does have this information.

Randomisation of treatments to experimental units will mean that treatment allocation is impartial – no treatment is deliberately favoured. However,

differences in the experimental units might favour one treatment over another and so introduce bias. So:

> Another aim in designing experiments is to remove or reduce potential sources of bias.

Thus, for example:

- In the mustard seed experiment the two pots were placed side by side (step 13 of the experiment) so that the two pots were at a similar temperature. For the same reason, you swapped the positions of the two large containers each day (Subsection 2.4 of Unit 10).
- The Clackmannanshire experiment examined three different methods of teaching children to read. Teaching methods are applied to whole classes, rather than individual children, so care was taken in selecting the participating classes and schools to ensure that the children taught by each method were broadly comparable (Subsection 1.3, Unit 8).
- Many experiments (notably many clinical trials) are group-comparative trials in which people/patients are allocated to treatments at random, but with restrictions, so that the overall characteristics of each treatment group are similar for qualities that are thought to matter. If gender and age were thought to influence a person's response to treatment, then care would be taken to ensure that the control group and treatment group (or treatment groups) contained similar male-to-female ratios and that each group had similar age profiles.

Activity 11 Group-comparative trial of an arthritis drug

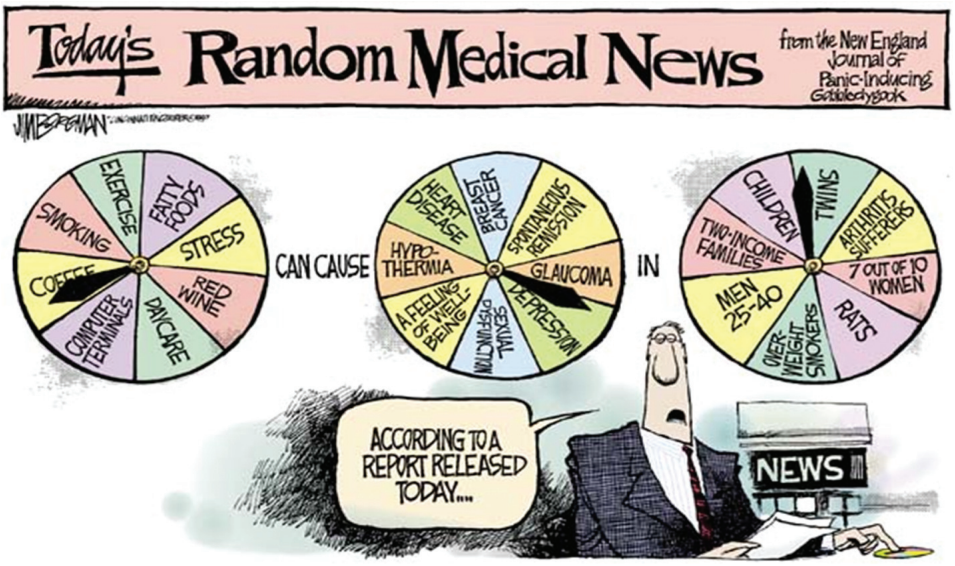
A drug company wishes to carry out a clinical trial on a new drug that, it is hoped, will alleviate the symptoms of arthritis. A design is chosen in which 20 arthritis sufferers are allocated to the trial. Half of the patients are to receive the new drug for four weeks (the experimental group), and the other half are to receive an existing drug for the same period of four weeks (the control group). The ten patients for the experimental group are chosen as a simple random sample from the list of 20. The remaining ten patients form the control group.

Once the allocation has been carried out, the research staff running the clinical trial discover to their dismay that all the patients allocated to the experimental group turn out to be women and all those allocated to the control group turn out to be men.

(a) Explain what characteristics of this experiment make it a group-comparative trial.

(b) Explain why the allocation of all the women to the experimental group and all the men to the control group might upset the validity of the clinical trial.

(c) If the research staff were to abandon this trial and start again with 20 new patients, how might they alter their allocation procedure to ensure that such a problem could not arise?



Random variation will affect differences that are observed between treatments, so a third aim when designing an experiment is to try to reduce this variation.

Quite commonly, an experiment is designed to yield pairs of closely related measurements: the difference between the measurements in each pair is calculated and these differences form the data that are used in testing hypotheses or forming confidence intervals. The differences typically have a much smaller variance than the original measurements. The following are instances from earlier units where differences between pairs were formed.

- In Exercise 9 of Unit 6 (Section 5), the change in degree of depression after taking methadone (before $-$ after) was the data used for a hypothesis test of whether methadone had any effect, rather than just the degree of depression after taking the drug.
- Matched-pairs t-tests are based on the differences within pairs of observations. A pair of measurements might be the weights of an object given by two different weighing machines (Activity 22 in Subsection 4.2 of Unit 10) or resting heart rate before and after an exercise program (Activity 26 in Section 6 of Unit 10), for example.
- In clinical trials, matched pairs look at differences between pairs of patients who are closely matched (twins ideally!). One treatment is given to one member of the pair and the other treatment to the other member. Often, though not always, the data will be analysed using a paired t-test.
- In a crossover design, each person taking part in the trial is given both treatments. Thus, each individual acts as their own control, thereby reducing variability. The differences between a person's responses under the two treatments are the data analysed.

When designing an experiment, a common question is: *How much data should be collected?* Often the experimenter is hoping to minimise the amount of data he or she has to collect. However, the experimenter should also consider the amount of time that will be spent preparing for the experiment, analysing the data statistically and then writing reports or papers about the results. These activities will take a significant amount of time – several weeks or months if the work is to be reported as a project or published in a scientific journal. In such cases, gathering data is just a small part of the process – but the quantity of data that is

gathered is often crucial to the value of an experiment. The chance of rejecting a null hypothesis that is false is dependent on the amount of data gathered. Hence:

> Good advice for when you are designing an experiment is that you should aim to gather as much data as you reasonably can.

Exercises on Section 2

Exercise 1 Stratified sampling and cluster sampling

(a) Consider again the population in Table 3 (Subsection 2.1). Suppose sets A and B can sensibly be considered as one stratum, sets C and D as a second stratum, and sets E and F as a third stratum. A stratified sample is required in which six people are selected from the first stratum, six from the second stratum, and five from the third stratum. (The third stratum is smaller than the others.) Start in row **36** of the random number table (Unit 4 appendix) and pick a stratified sample. Give the labels of the people in the sample.

(b) Suppose that the sets in Table 3 are groups of people who are widely separated geographically, so that cluster sampling is the obvious survey method to use. Restricting your sample to just two of the sets, sample approximately one-third of the individuals in each cluster.

Obtain the sample using the following procedure:

- Label the sets A–F from 1 to 6. Using *single* random digits, and starting at the beginning of row **95** of the random number table, select the two sets to be sampled. These sets are to be sampled in the order in which they are selected.
- Determine the sizes of the samples to take from each cluster by dividing each cluster size by 3 and rounding the results *up* to whole numbers.
- To select individuals for the subsample from the first selected street, use pairs of digits starting at row **72** of the random number table. No person may be selected more than once. To select individuals from the second subsample, continue from the point reached in the random number table after selecting the first subsample, and apply the same procedure again.

List the people chosen in the subsamples.

Exercise 2 Survey of NHS trust employees

A large NHS trust wishes to survey a sample of its 6000 employees about job satisfaction. In addition to each employee's payroll number, the employee database also includes information about each employee's age and grade. It is thought that both age and grade might relate to job satisfaction.

(a) Describe how the procedures of stratified sampling and systematic sampling might together be used to choose a sample of size 500 that represents an appropriate balance of the workforce.

(b) Identify two non-sampling errors that may arise in the survey.

Exercise 3 Other trials of an arthritis drug

Suppose that, as in Activity 11 (Subsection 2.2), a drug company wishes to carry out a clinical trial to compare the usefulness of two drugs (a new drug and an existing drug) for alleviating the symptoms of arthritis.

(a) Describe briefly how researchers could use a crossover design in a clinical trial comparing the two drugs.

(b) Describe briefly how researchers could use a matched-pairs design in a clinical trial comparing the two drugs.

(c) Which of the two designs would you consider the more suitable? Would a group-comparative trial be better?

3 Probability

In this section, we first review some properties of probability that were given in Units 6 and 8. We then consider the binomial distribution and the probabilities that it gives. You used a specialised form of the binomial distribution to test hypotheses about the median in Unit 6, and here the general form of the distribution is described.

3.1 Basic properties

The data in Table 4 will be used to illustrate the basic properties of probability. It concerns academic statisticians working in UK universities in 2013. They are separated into five age-bands (under 30, 30–39, 40–49, 50–59, and 60 or over) and the table gives the number of them in each job category and age-band.

Table 4 Academic statisticians in UK universities

	< 30	30–39	40–49	50–59	≥ 60	Total
Research Fellow	102	82	31	5	7	227
Lecturer	13	150	59	11	6	239
Senior Lecturer	0	35	70	43	8	156
Professor	0	8	51	66	47	172
Total	115	275	211	125	68	794

'Research fellow' includes research assistants as well as research fellows, and 'senior lecturer' includes readers as well as senior lecturers.

Definition of probability

A simple definition of probability is to equate it to a *proportion*:

$$\text{Probability} = \text{Proportion}.$$

More precisely:

Probability of an event

Let E stand for the event of selecting a person or object with some particular property from a population using random sampling. Then the

probability of E is given by

$$P(E) = \frac{\text{Number in population with particular property}}{\text{Total number in population}}.$$

Example 1 Probability of picking a lecturer

There are 239 lecturers in the population of 794 academic statisticians in Table 4. Thus, the proportion of lecturers in the population is

$$\frac{239}{794} \simeq 0.301.$$

Hence, if we pick a person at random from the academic statisticians, the probability that we pick a lecturer is 0.301.

Activity 12 Picking a research fellow

Suppose an academic statistician is picked at random. What is the probability that the person is:

(a) a research fellow?

(b) aged 30–39?

(c) a research fellow aged 30–39?

Prominent statisticians from the University of Cambridge Statistical Laboratory, 1953

The figure below shows staff and postgraduate students of the University of Cambridge Statistical Laboratory in 1953.

Some of the people shown were already very well known statisticians in 1953, and others became prominent statisticians later. You will hear of some of them if you study statistics further, including David Cox (now Sir David Cox, FRS, FBA), John Wishart (first Director of the Statistical Laboratory), Frank Anscombe and Dennis Lindley, who are respectively third to sixth from the left in the second row.

Conditional probability

Sometimes we want probabilities for subpopulations. For example, Smith (a statistician) has become a professor at the age of 37 and wants to know the probability that an academic statistician is a professor if they are in the 30–39 age-band. Thus academic statisticians aged 30–39 form the subpopulation of interest.

From Table 4, there are 275 in this subpopulation, of whom 8 are professors. Hence among statisticians aged 30–39, the proportion who are professors is

$$\frac{8}{275} \simeq 0.029.$$

Thus

$$P(\text{professor}|\text{aged 30–39}) \simeq 0.029.$$

The conditional probability of A, given B, denoted $P(A|B)$, is the probability that A occurs, given that B occurs.

Activity 13 Senior lecturer aged 40–49 years old

Suppose an academic statistician is picked at random.

(a) If the person is aged 40–49, what is the probability that they are a senior lecturer?

(b) If the person is a senior lecturer, what is the probability that they are aged 40–49?

Independence

Independence is an important concept in statistics. Most of the hypothesis tests in M140 require observations to be independent. In particular, this is true of the sign test, the one-sample and two-sample z-tests, and t-tests for one sample or two unrelated samples. In such contexts, the everyday notion of independence corresponds to the statistical definition. We are assuming that the value taken by one (random) observation has no influence on the value taken by any other random observation. That is, as in standard English, two things are 'independent' if they are unrelated.

More precisely, we define statistical independence in terms of probabilities.

Hoping to avoid the difficulties of using conditional probability, Thomas Jefferson writes the Declaration of Independence.

Two events A and B are statistically independent if the occurrence of one has no influence on the chance of occurrence of the other. Then

$$P(A|B) = P(A).$$

[It can be shown that if $P(A|B) = P(A)$, then $P(B|A) = P(B)$.]

Events that are unrelated are also statistically independent. For example, if you roll a die and toss a coin, the event 'the die gives a three' is both physically independent and statistically independent of the event 'the coin lands heads'. However, statistical independence is a numerical property of probability, and events can be statistically independent without being physically disconnected. This is illustrated in the following example.

Example 2 Rolling a die

Suppose an ordinary six-sided die is rolled. Assuming it is unbiased, it is equally likely to roll a 1, 2, 3, 4, 5 or 6. Suppose it is rolled once and consider the following events,

- A: it rolls an even number (i.e. 2, 4 or 6).
- B: it rolls a 4 or more (i.e. 4, 5 or 6).
- C: it rolls a 3 or more (i.e. 3, 4, 5 or 6).

Which of these events are independent? Well,

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

as three of the six possibilities result in A occurring.

Now if we know that B has occurred, then our population of possible outcomes reduces to 4, 5 or 6. Each of these three events is equally likely, and event A occurs if the roll was actually a 4 or 6 (but not if it was a 5). Hence,

$$P(A|B) = \frac{2}{3}.$$

As $P(A)$ does not equal $P(A|B)$, the probability that A occurs is affected by whether or not B occurs. Thus events A and B are not independent.

Suppose, instead, that we know that C has occurred. Then our population of possible outcomes is 3, 4, 5 or 6. Each of these four outcomes is equally likely, and event A occurs if the roll was a 4 or 6 (but not if it was a 3 or 5). Hence,

$$P(A|C) = \frac{2}{4} = \frac{1}{2}.$$

As $P(A)$ does equal $P(A|C)$, the occurrence of C has no influence on the probability that A occurs. Thus events A and C *are* independent, even though the same physical quantity – the outcome of rolling a die – determines them both.

Lastly, for events B and C,

$$P(C) = \frac{4}{6} = \frac{2}{3}$$

as four of the six possibilities result in C occurring. Also, if we know that B has occurred, then our population of possible outcomes is 4, 5 or 6. Regardless of which of these outcomes has occurred, C will occur. That is, C is certain to occur if B occurs, so

$$P(C|B) = 1.$$

As $P(C)$ does not equal $P(C|B)$, the probability that C occurs is affected by whether or not B occurs. Thus events B and C are not independent.

Joint probabilities (the 'and' linkage)

The joint probability of A and B, denoted $P(A \text{ and } B)$, is the probability that both A and B occur together.

Joint probabilities

Let A and B be any two events. Joint and conditional probabilities are linked by the following relationships:

$$P(A \text{ and } B) = P(A) \times P(B|A) = P(B) \times P(A|B).$$

Also, if A and B are independent events [so $P(B|A) = P(B)$], then

$$P(A \text{ and } B) = P(A) \times P(B).$$

Activity 14 Picking a professor

Suppose an academic statistician is picked at random.

(a) How many academic statisticians are professors aged 50–59? Hence show that 0.083 is the probability that the randomly picked person is a professor aged 50–59.

(b) What is the probability that the randomly picked person is a professor?

(c) If the randomly picked person is a professor, what is the probability that they are aged 50–59?

(d) Define events A and B as follows.

A: a randomly picked academic statistician is a professor

B: a randomly picked professor is aged 50–59.

Use (a), (b) and (c) to show that

$$P(A \text{ and } B) = P(A) \times P(B|A).$$

Adding probabilities (the 'or' linkage)

We say that the event 'A or B' occurs if (i) A occurs, or (ii) B occurs, or (iii) *both* A *and* B occur. The 'or' linkage leads to the addition of probabilities and has a simpler form when events are mutually exclusive.

Two events are said to be mutually exclusive if they cannot occur at the same time. More generally, any number of events are said to be mutually exclusive if no two of them can occur at the same time.

If events A and B are mutually exclusive, then $P(A \text{ and } B) = 0$, as they cannot both occur at the same time.

Addition rules for probabilities

Let A and B denote two events, mutually exclusive or not. Then

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

If A and B are mutually exclusive,

$$P(A \text{ or } B) = P(A) + P(B).$$

Activity 15 The great and the good

(a) Suppose a statistical organisation wishes to invite 'the great and the good' to an event it is hosting. Any UK academic statistician who is a professor or is aged at least 60 will be invited. Using Table 4, explain why 193 UK academic statisticians will be invited. How many people are double-counted if you simply add the number of professors to the number of academic statisticians aged 60 or over?

(b) At a more select gathering, only older professors are invited (professors aged at least 60). How many people from UK universities are invited?

(c) Define events A and B as follows.

A: a randomly picked academic statistician is a professor

B: a randomly picked academic statistician is aged 60 or over.

Use (a) and (b) to calculate the probabilities $P(A \text{ or } B)$ and $P(A \text{ and } B)$.

(d) Also calculate $P(A)$ and $P(B)$. Hence show that

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B),$$

in line with theory.

We have focused on the case where there are exactly two events. The following box gives useful results for the case where there are four events, which can be easily generalised to other numbers of events.

Sets of events

If A, B, C and D are independent events, then

$$P(A \text{ and } B \text{ and } C \text{ and } D) = P(A) \times P(B) \times P(C) \times P(D).$$

If A, B, C and D are mutually exclusive events, then

$$P(A \text{ or } B \text{ or } C \text{ or } D) = P(A) + P(B) + P(C) + P(D).$$

3.2 The binomial distribution

This subsection contains new material, not previously covered in M140.

You encountered the binomial distribution in Unit 6. It is an important distribution in statistics so here we consider it further.

In Unit 6 (just before Activity 22 in Subsection 3.1), we determined that the number of ways in which a committee of 3 people can be chosen from 10 members of a club is equal to

$$\frac{10 \times 9 \times 8}{3 \times 2 \times 1}.$$

This led to the following result.

Number of combinations

Suppose there are n objects to choose from. Then the number of ways of choosing x objects if the order does not matter is

$$^nC_x = \frac{n \times (n-1) \times \cdots \times (n-x+1)}{x \times (x-1) \times \cdots \times 1}.$$

(There are x terms in both $n \times (n-1) \times \cdots \times (n-x+1)$ and $x \times (x-1) \times \cdots \times 1$.)

For any value of n, $^nC_0$ and ^nC_n are defined as being equal to 1.

The result will be used to obtain the probabilities given by a binomial distribution. The following activity will refresh your memory of using it.

Activity 16 Socks for the weekend

(a) You have six clean pairs of socks and must pack three pairs in your bag to go away for the weekend. In how many ways could you choose the three pairs to take?

(b) A month later you are going away for a longer break and must pack four pairs of socks. If you have nine clean pairs, how many choices do you have? (Either you have got better at doing the washing or you have bought some more socks!)

Suppose that a defence lawyer has a success rate of 0.3. That is, in 30% of the trials in which he is the defence lawyer the defendant is acquitted. Let S denote success (the defendant is acquitted) and F denote failure (the defendant is found guilty). Then

$$P(S) = 0.3 \quad \text{and} \quad P(F) = 1 - 0.3 = 0.7.$$

A sequence of five trials might give the sequence $SFFSS$, where this means that the first trial was a success for the lawyer, the next two were failures, and the last two were successes. If we assume that outcomes are independent of each other, then

$$P(SFFSS) = 0.3 \times 0.7 \times 0.7 \times 0.3 \times 0.3 = 0.3^3 \times 0.7^2.$$

Other sequences that gives 3 successes and 2 failures include $SSSFF$ and $FSSSF$, with probabilities

$$P(SSSFF) = 0.3 \times 0.3 \times 0.3 \times 0.7 \times 0.7 = 0.3^3 \times 0.7^2$$

and

$$P(FSSSF) = 0.7 \times 0.3 \times 0.3 \times 0.3 \times 0.7 = 0.3^3 \times 0.7^2.$$

Clearly, the probability of any sequence that gives 3 successes and 2 failures is $0.3^3 \times 0.7^2$. Suppose now, that we want the probability that exactly 3 of the next 5 trials will be a success (in any order). It follows that this probability equals:

$$\text{(number of sequences of 5 trials giving 3 successes)} \times 0.3^3 \times 0.7^2.$$

As ${}^5C_3$ is the number of sequences of 5 trials giving 3 successes,

$$\begin{aligned} P(\text{3 successes in 5 trials}) &= {}^5C_3 \times 0.3^3 \times 0.7^2 \\ &= \frac{5 \times 4 \times 3}{3 \times 2 \times 1} \times 0.027 \times 0.49 \\ &= 0.1323. \end{aligned}$$

Activity 17 Sales targets

A salesman has a daily target for the number of sales he should make in a day. Let S denote success (he reaches his target) and F denote failure. Assume that the probability of S on any day is 0.6 (so $P(F) = 0.4$) and that results on different days are independent of each other.

(a) What is the probability that in the next 6 days his sequence of successes and failures is $FSSSFS$?

(b) How many different sequences of 6 days give 4 successes and 2 failures?

(c) What is the probability that the salesman meets his target in exactly 4 of the next 6 days?

(d) What is the probability that the salesman meets his target in exactly 5 of the next 8 days?

The probabilities you calculated in parts (c) and (d) of Activity 17 are probabilities from a **binomial distribution**. More generally, the binomial distribution arises in the following situation.

1. There is a fixed number of trials, where a trial is an event that can result in success (S) or failure (F). Let n denote the number of trials.
2. The probability of success in a trial is the same for all trials and the result of any trial is independent of the results of other trials. Let p denote the probability of success and $q = 1 - p$ denote the probability of failure.
3. Let x denote the number of successes in the n trials. Then the probability distribution of x is a binomial distribution.

In the example with lawyers a 'trial' was actually a real trial. We were interested in the number of successes in 5 trials, so n equals 5, and the probability of success in any one trial was 0.3, so $p = 0.3$ and $q = 0.7$.

In the example with the salesman in Activity 17, a 'trial' was a day's sales performance and success equated to making the daily target. Thus $p = 0.6$ was the probability of success and $q = 0.4$ was $P(F)$. In part (c) of Activity 17, we were concerned with a 6-day period, so n was 6, and we wanted $P(\text{4 successes}) = P(x = 4)$. In (d), n was 8 and we wanted $P(x = 5)$.

The following defines the binomial distribution.

Binomial distribution

Suppose the result of a trial is success or failure (with no other possibilities). Suppose also that the probability of success (p) is the same in each trial. Put $q = 1 - p$ and let x denote the number of successes in n trials. If trials are independent of each other, then

$$P(x) = {}^nC_x \times p^x \times q^{n-x}.$$

The binomial distribution is the probability distribution of x.

Activity 18 Blood group O

Forty-four per cent of the population of the UK are blood group O. If seven people are picked at random, what is the probability that exactly two of them are blood group O?

In Unit 6, we were interested in the number of observations in a sample that would exceed the population median. Consider each observation as a trial and equate success to 'the observation exceeds the population median'. Then, if there are x successes and n observations,

$$P(x) = {}^nC_x \times p^x \times q^{n-x}.$$

From the definition of the median, $p = P(\text{success}) = \frac{1}{2}$ and $q = 1 - p = \frac{1}{2}$. Thus the probability that there are x values above the population median in a sample of n observations is

$${}^nC_x \times \left(\frac{1}{2}\right)^x \times \left(\frac{1}{2}\right)^{n-x} = {}^nC_x \times \left(\frac{1}{2}\right)^n.$$

This formula (the formula for a binomial probability when $p = \frac{1}{2}$) is the special case of the binomial distribution used in Unit 6.

You have now covered the material related to Screencast 2 for Unit 12 (see the M140 website).

You have also now covered the material needed for Subsection 12.1 of the Computer Book.

Exercises on Section 3

Exercise 4 Age and job category independent?

Consider Table 4 (Subsection 3.1) and suppose an academic statistician is picked at random.

(a) Are the events 'the person is aged 40–49' and 'the person is a lecturer' independent?

(b) Use the formula $P(A \text{ and } B) = P(B) \times P(A|B)$ to obtain the probability that the person is a lecturer aged 40–49.

(c) What is the probability that the person is a lecturer *or* aged 40–49 *or* both?

Exercise 5 Binomial probabilities

Suppose the probability of success in a trial is 0.2 and that trials are independent of each other.

(a) If there are four trials, what is the probability that there is exactly one success?

(b) If there are five trials, what is the probability that there are exactly two successes?

(c) If there are seven trials, what is the probability that there are no successes?

Exercise 6 Relief from headache

The probability that a person with a headache will feel better within an hour of taking a particular headache tablet is 0.6. If six people with headaches take the tablet, determine the following probabilities, stating any assumptions you make.

(a) The probability that exactly two of them feel better within an hour.

(b) The probability that two or fewer of them feel better within an hour.

4 Hypothesis testing and contingency tables

Hypothesis testing is an important use of statistics. In this section we review the structure of a hypothesis test, illustrating the main ideas by describing the χ^2 test for contingency tables.

The following is the procedure in a hypothesis test.

1. The first step is to state the null hypothesis, H_0, and the alternative hypothesis, H_1. We provisionally assume that H_0 is true and (usually) hope to find that the data cast strong doubt on this assumption. If that happens, then we will have evidence against the null hypothesis.
2. From the data, we construct a test statistic whose value relates to the null hypothesis. Often the value of this statistic should be small if H_0 is true, so that a large value casts doubt on H_0.
3. Lots of different test statistics may relate to H_0. We choose one whose probability distribution is fully known if H_0 is true. For example: in Unit 7, we used test statistics that followed a normal distribution if H_0 was true; in Unit 8,

test statistics followed a χ^2 distribution if H_0 was true; in Unit 10, most of the test statistics followed a t distribution if H_0 was true; while in the sign test in Unit 6, the test statistic follows a binomial distribution with $p = q = \frac{1}{2}$ if H_0 is true.

4. The possible values that the test statistic might have taken are divided into two sets:
   - Set A contains those values that are at least as unlikely as the value given by the sample data, if H_0 holds.
   - Set B contains values that are relatively likely to occur if H_0 is true. Specifically, it contains those values that are more likely to occur, if H_0 is true, than the actual value given by the sample data.
5. The actual value taken by the test statistic is in set A. We determine the probability of observing a value from that set. This probability is the p-value from the test. If it is small then either a very unusual event has happened, or H_0 is incorrect.
6. The p-value is interpreted as follows:

$p > 0.10$	Little evidence against H_0
$0.10 \geq p > 0.05$	Weak evidence against H_0
$0.05 \geq p > 0.01$	Moderate evidence against H_0
$0.01 \geq p > 0.001$	Strong evidence against H_0
$0.001 \geq p$	Very strong evidence against H_0

   Unless we are using Minitab (or some other statistical software) we will seldom know the exact p-value. Instead the test statistic is compared with *critical values* given in an appropriate table (such as Table 28 in Subsection 4.4 of Unit 8 for χ^2 tests, or Table 2 in Subsection 3.3 of Unit 10 for t-tests – versions of both are in the Handbook). This determines the significance level at which H_0 can be rejected.
7. The conclusion from the hypothesis test is summarised in plain English. For example, the conclusion might be: *There is strong evidence that the new method is better than the old method.*

We next consider the hypothesis test of independence between the row and column variables of a contingency table, using the following example.

Example 3 Coffee consumption

A researcher wanted to examine whether a person's coffee consumption was associated with their age. She selected a random sample of 180 people and classified each person according to whether their age was under 35, 35–60 or at least 61, and whether their coffee consumption was low, medium or high. Results are given in Table 5. (The data are artificial.)

Table 5 Level of coffee consumption by age group

	Coffee consumption			
	Low	Medium	High	Total
Under 35	24	41	25	90
35–60	4	32	24	60
61 and over	8	17	5	30
Total	36	90	54	180


Coffee consumption by age group

The researcher was interested in whether there is an association between age and the level of coffee consumption, or whether these variables are independent. When examining these questions using the data in a contingency table, the null hypothesis is that the row and column variables are independent. This assumption of independence enables the Expected value in each cell to be calculated.

The alternative hypothesis is that row and column variables are not independent. Making the assumption 'row and column variables are not independent' would not enable us to calculate the Expected values in the cells, so that must not be chosen as the null hypothesis.

In this example the hypotheses are:

H_0: Age group and coffee consumption are independent.

H_1: Age group and coffee consumption are not independent.

For the hypothesis test, Expected values are needed. If the variables are independent and we treat the row totals and column totals as fixed, what values might you expect for the rest of Table 5? That is, suppose we knew the values in Table 6 (and nothing else), what would be reasonable estimates of other values in the table?

Table 6 Row and column totals

	Coffee consumption			
	Low	Medium	High	Total
Under 35				90
35–60				60
61 and over				30
Total	36	90	54	180

Well, half the people in the sample are aged under 35, so we would expect half of the people with a low level of coffee consumption to be under 35, half the people with medium coffee consumption to be under 35, and half the people with high coffee consumption to be under 35. Hence the Expected values for the top row of the table are $36/2 = 18$, $90/2 = 45$ and $54/2 = 27$.

Similarly, one-third of the people in the sample are aged 35–60, so we would expect one-third of the people with a low level of coffee consumption to be in that age group ($36/3 = 12$), one-third of the people with medium coffee consumption to be in that age group ($90/3 = 30$), and one-third of the people with high coffee consumption to be in that age group ($54/3 = 18$). For the '61 and over' group the Expected values are one-sixth of 36, 90 and 54 for the three levels of coffee consumption (i.e. 6, 15 and 9).

These values form the Expected table:

	Low	Medium	High	Total
Under 35	18	45	27	90
35–60	12	30	18	60
61 and over	6	15	9	30
Total	36	90	54	180

The general formula for calculating the Expected values is usually written as:

$$\text{Expected value} = \frac{\text{Row total} \times \text{Column total}}{\text{Overall total}}.$$

Thus, the Expected value for the first cell would usually be obtained from

$$\frac{90 \times 36}{180} = 18$$

and similarly for the other cells. (Check that the Expected values are all greater than or equal to 5, so that the χ^2 test is valid.)

Why might the data cast doubt on H_0? Well, the values in the Expected table, which were calculated under the assumption that H_0 is true, are not equal to the values observed in the sample, given in Table 5. The differences are the residuals. That is,

$$\text{Residual} = \text{Observed} - \text{Expected}.$$

The Residual for the first cell is $24 - 18 = 6$ and the following is the complete Residual table.

	Low	Medium	High
Under 35	6	-4	-2
35–60	-8	2	6
61 and over	2	2	-4

Now, even if row and column variables are independent (so that H_0 holds), the residuals would seldom all equal 0, because sample data is affected by random variation. The question is whether the residuals are so large that we should reject H_0. To answer this question, we must combine the information provided by the different residuals to form a test statistic. The distribution of this statistic must be fully known when H_0 is true. To this end, we calculate the χ^2 contribution of each cell and add these together. A cell's χ^2 contribution is

$$\frac{(\text{Residual})^2}{\text{Expected}}.$$

So the χ^2 contribution of the first cell is $6^2/18 = 2$.

The complete table of χ^2 contributions is:

	Low	Medium	High
Under 35	2	0.3556	0.1481
35–60	5.3333	0.1333	2
61 and over	0.6667	0.2667	1.7778

The χ^2 test statistic is the sum of the nine χ^2 contributions:

$$\begin{aligned}\chi^2 &= 2 + 0.3556 + 0.1481 + 5.3333 + 0.1333 + 2 + 0.6667 \\ &\quad + 0.2667 + 1.7778 \\ &\simeq 12.682.\end{aligned}$$

If H_0 is true, then the test statistic follows a χ^2 distribution. For an Observed table with r rows and c columns,

$$\text{degrees of freedom} = (r-1) \times (c-1).$$

Here the Observed table is a 3×3 table, so its degrees of freedom are $(3-1) \times (3-1) = 4$. Hence we compare the test statistic (12.68) with a χ^2 distribution on 4 degrees of freedom. From Table 28 in Subsection 4.4 of Unit 8 (and in the Handbook), the 5% and 1% critical values are CV5 $= 9.488$ and CV1 $= 13.277$.

Since $12.682 > 9.488$, we reject H_0 at the 5% significance level but, as $12.682 < 13.277$, we do not reject H_0 at the 1% significance level.

We conclude that there is moderate evidence that the level of coffee consumption varies with age, but the evidence is not strong.

Examination of the χ^2 contributions shows that the biggest value is for the low level of coffee consumption in the 35–60 age range. The residual for this cell is negative. That is, the number of 35- to 60-year-olds having a low level of coffee consumption is smaller than expected under H_0. This is the main sample evidence of departure from independence.

Activity 19 Alternative test statistic?

To decide whether the Residuals are so large that we should reject H_0, we formed the test statistic

$$\sum \frac{(\text{Residual})^2}{\text{Expected}}.$$

Suggest a reason for not using the simpler quantity

$$\sum (\text{Residual})^2$$

as the test statistic.

Exercises on Section 4

Exercise 7 Study of air pollution

In a study of air pollution, a random sample of 100 households was selected in each of four localities. Each householder was asked if one or more members in the household were concerned by the level of air pollution. A summary of the responses is given in Table 7.

Table 7 Households with a concern about air pollution

	One or more members concerned		
Locality	Yes	No	Total
A	25	75	100
B	16	84	100
C	12	88	100
D	30	70	100
Total	83	317	400

A hypothesis test is required of whether concern about air pollution varies with locality.

(a) Write down the null and alternative hypotheses.

(b) Obtain the Expected table.

(c) Hence obtain the Residual and χ^2 contributions tables.

(d) Calculate the χ^2 test statistic and note the appropriate critical values, CV5 and CV1 (using Table 28 in Subsection 4.4 of Unit 8 and in the Handbook).

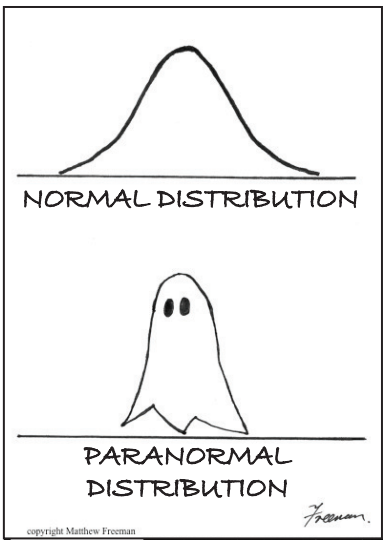
(e) State your conclusions.

5 z-tests, t-tests and confidence intervals

Testing hypotheses about a population mean and forming a confidence interval for a population mean are common statistical tasks. So are testing hypotheses about the difference between the means of two populations and forming a confidence interval for that difference. These tasks are discussed in Subsection 5.2. Before that, in Subsection 5.1, underpinning results related to a normal distribution are reviewed.

5.1 The normal distribution

A normal distribution that has a mean $\mu = 5$ and a standard deviation $\sigma = 2$ is shown in Figure 5(a). The distribution is bell-shaped and, as it is symmetric, the median and mode also equal 5. Figure 5(b) shows the standard normal distribution, which has mean $\mu = 0$ and standard deviation $\sigma = 1$.



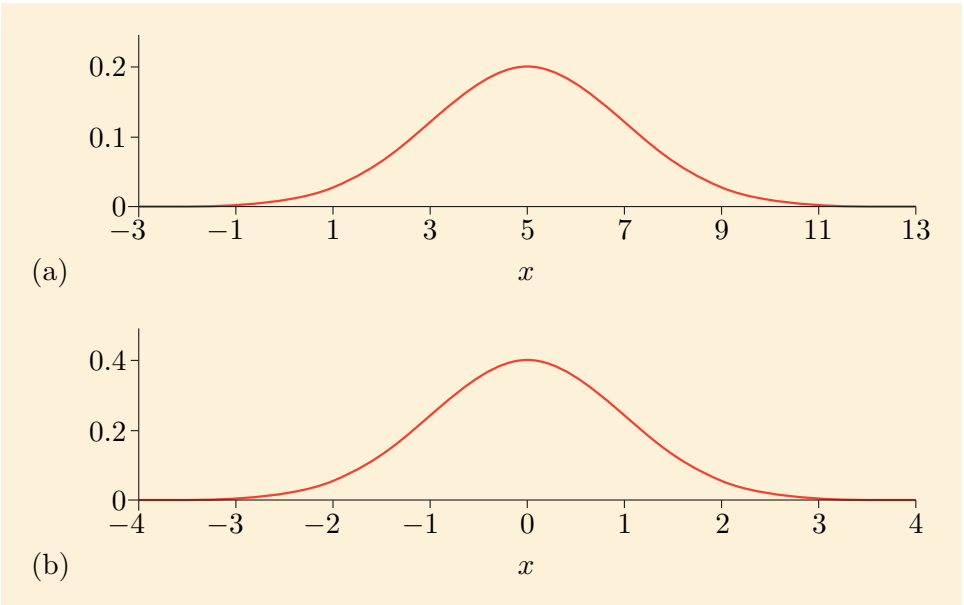


Figure 5 (a) A normal distribution with mean 5 and standard deviation 2; (b) a standard normal distribution.

Comparison of the two figures illustrates that all normal distributions have identical shapes. Consequently, we can transform from any normal distribution to the standard normal distribution.

Transforming a normal distribution to the standard normal distribution

If a variable x has a normal distribution with mean μ and standard deviation σ, then the variable

$$z = \frac{x - \mu}{\sigma}$$

has the standard normal distribution.

Activity 20 Shell thicknesses

The shell thickness, x, of eggs produced by a large flock of White Leghorn hens follows (approximately) a normal distribution with mean $\mu = 0.38$ mm and standard deviation $\sigma = 0.03$ mm.

Calculate the value of z corresponding to each of the following values of x (in mm). In each case, interpret your answer by completing a sentence of the form 'So a shell thickness of *** mm is *** standard deviations *** than the mean thickness of *** mm.

(a) $x = 0.40$; (b) $x = 0.30$.

As noted in Activity 20, the thicknesses of eggshells of White Leghorn hens approximately follow a normal distribution. The heights of men in Scotland also approximately follow a normal distribution (as noted in Subsection 3.1 of Unit 7). So do the heights of 7-year-old boys, the cholesterol levels of young adults, the weight gains of calves on a standard diet and many other quantities. One reason that the normal distribution is important is because many natural quantities follow approximately a normal distribution. Indeed, the normal distribution is so ubiquitous that it is common to assume that a quantity follows a normal distribution unless there is reason to think otherwise. Thus, in the experiment where you grew mustard seedlings, you assumed that root length was normally distributed, both for seedlings grown in the light and those grown in the dark.

White Leghorn hens

Of course, you should not assume that a quantity follows a normal distribution if the data suggest otherwise, or if the situation suggests that the distribution will not be normal. So, for instance, it would be unwise to assume that incomes in the general population follow a normal distribution, because we know, from Unit 3 (Subsection 1.4), that distributions of incomes are usually right-skew, with a few people earning far more than the median income.

You have also met another major reason for the importance of the normal distribution: the sampling distribution of a mean is approximately normal, regardless of whether or not the distribution of individual observations is normal. This result is part of the *central limit theorem*. The central limit theorem also specifies the mean and variance of the distribution of the mean.

Approximate normality of the sampling distribution of the mean (central limit theorem)

If n is large, no matter what shape the population distribution, the sampling distribution of the mean for samples of size n will be approximately normal.

The mean will equal the population mean μ and the standard deviation will equal the standard error SE $= \sigma/\sqrt{n}$.

Figure 6 reproduces Figure 2 from Unit 7 (Section 2). It shows the proportions of students obtaining each examination mark in MS221 in one presentation.

The distribution is clearly not bell-shaped, so it is far from normal.

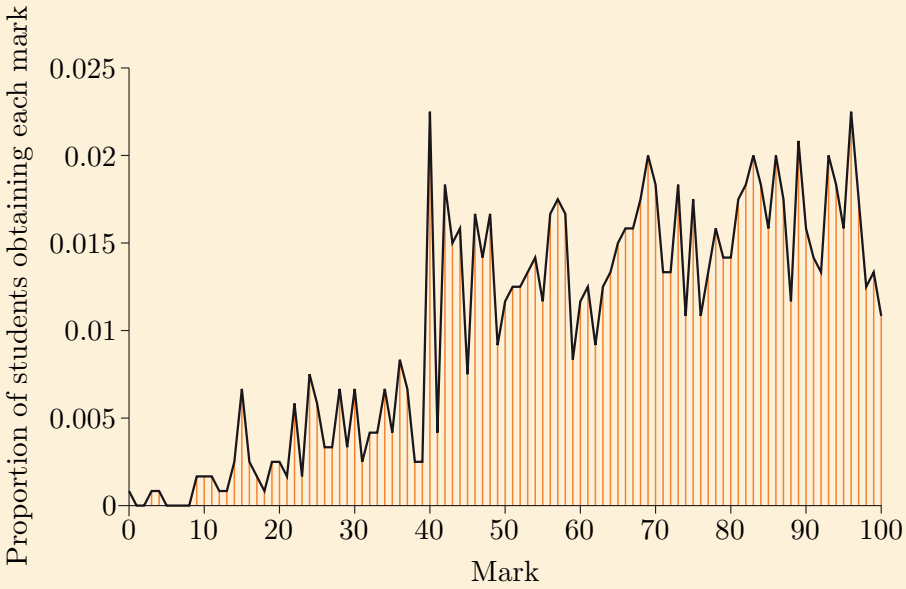


Figure 6 The distribution of MS221 exam marks

Suppose now that we took two students at random from the MS221 cohort and determined the mean of their two marks. The value of the mean depends on which two students we pick – repeatedly picking two students at random and calculating their mean mark will yield lots of different values. The distribution of the mean mark (from a sample of two students) is shown in Figure 7. The distribution is not bell-shaped, but is much closer to being bell-shaped than the distribution in Figure 6. If we take samples of 20 students (rather than just two) then the mean of their marks is normally distributed, approximately. That distribution is shown in Figure 8.

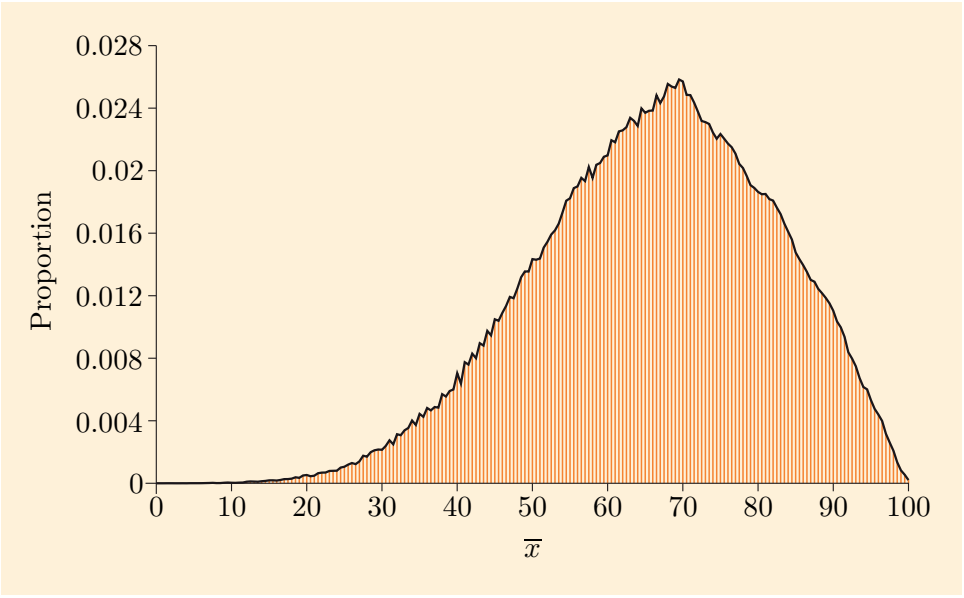


Figure 7 The distribution of the mean of 2 students' marks

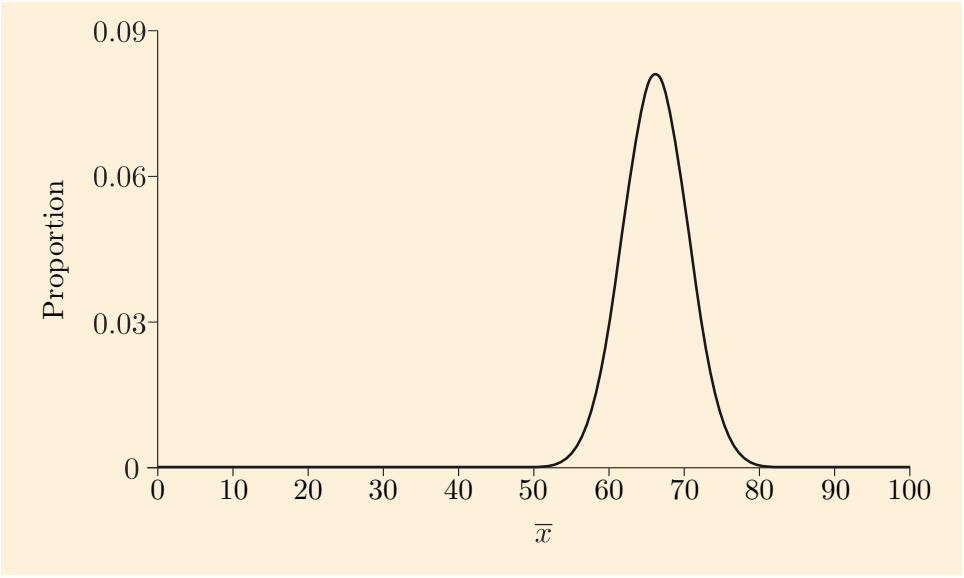


Figure 8 The distribution of the mean of 20 students' marks

Activity 21 Distributions of sample means

The distribution of the MS221 examination marks has mean $\mu = 66$ and standard deviation $\sigma = 22$.

(a) Give the mean and standard deviation of the sampling distribution of the mean for samples of size 5, and then for samples of size 15.

(b) Would the sampling distributions in (a) be approximately normal? Which would be closer to a normal distribution?

The Morane–Saulnier MS221 aircraft from 1928

5.2 Inference about the means of populations

This subsection contains new material, not previously covered in M140.

There are a number of different hypothesis tests for examining whether a population mean takes a specified value, or whether two populations have means that are equal; they have close similarities. We first consider how to choose the appropriate hypothesis test. A unified description of the tests is then given that highlights their similarities. A new hypothesis test is also introduced. Associated with each hypothesis test is a method of forming a confidence interval and these methods are described later in the section.

Which hypothesis test should be used?

M140 contains a number of different hypothesis tests for answering the following questions:

- Does the mean of a population take a specified value?
- Do the means of two populations have the same value?

For the first question, we have a single sample and would use one of the following tests:

- Test 1. One-sample z-test
- Test 2. One-sample t-test

For the second question, we have two samples of data, one from each of two populations. We would use one of the following tests:

- Test 3. Two-sample z-test
- Test 4. Matched-pairs t-test
- Test 5. Unpaired t-test for populations with a common variance
- Test 6. Unpaired t-test for populations with unequal variances

These tests make varied assumptions about observations following normal distributions and being random. Test 6 has not been covered in earlier units, but will be described later in this section. You will also learn in the Computer Book how to use Minitab to perform this test.

The flow chart in Figure 9 gives the steps to follow in order to decide which test should be used when there is just one population.

When there are two populations and we have a sample from each, one of the tests 3, 4, 5 or 6 would be used to test whether the population means are equal. The flow chart in Figure 10 gives the steps to follow in order to decide which of the tests to use (assuming one of these tests is suitable). The choice between test 5 or 6 depends upon whether ‘yes’ or ‘no’ is the response to the question ‘Population variances equal?’. Using the rule of thumb given in Unit 10 (Subsection 3.3), we treat the population variances as equal if the sample variances differ by a factor of less than three.

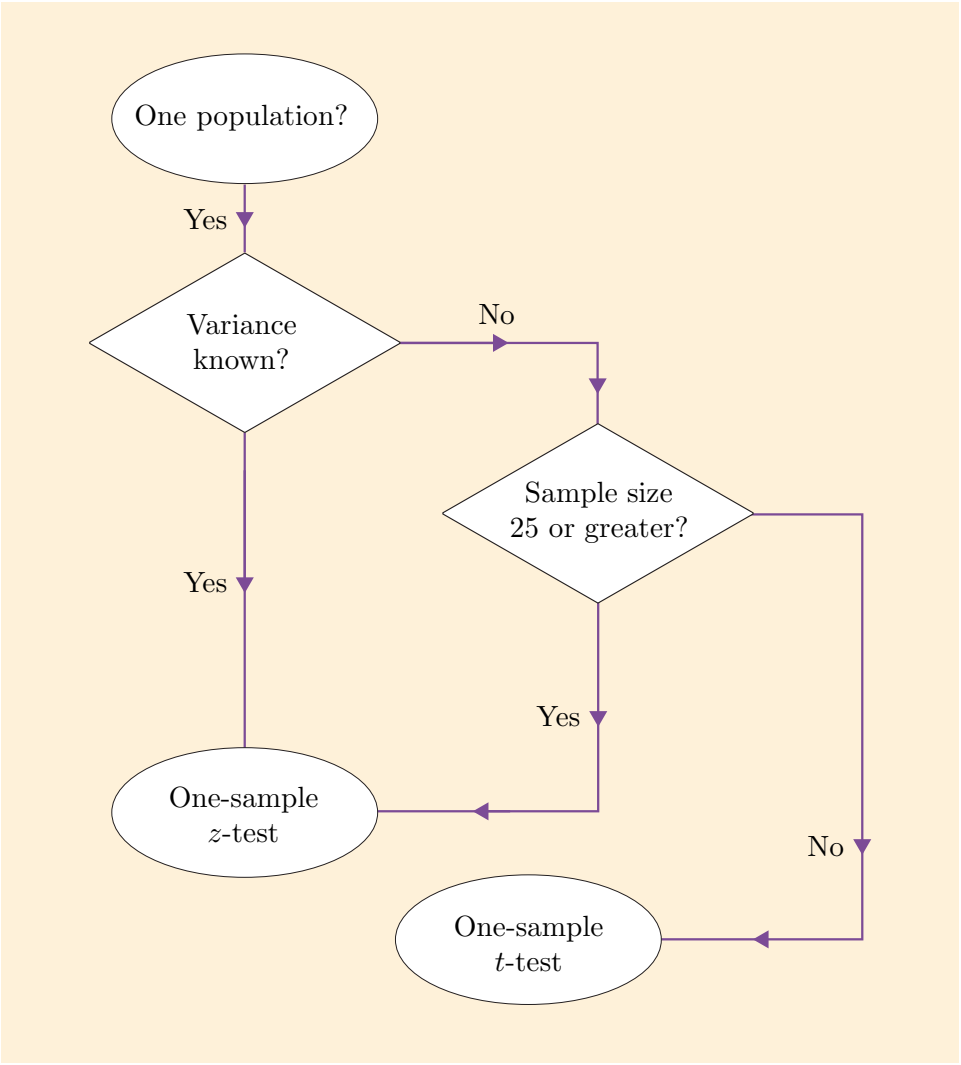


Figure 9 Flow chart for choosing a hypothesis test for inference about a population mean. (It is assumed that observations are random and that the population distribution is approximately normal if the sample size is small.)

Activity 22 Choosing a one-sample test

A sample is taken from a population in order to test the hypothesis that the population mean equals 15. What is the appropriate test for each of the following situations, assuming the population distribution is approximately normal?

(a) The sample size is 30 and the population variance is known.

(b) The sample size is 35 and the population variance is unknown.

(c) The sample size is 15 and the population variance is unknown.

(d) The sample size is 12 and the population variance is known.

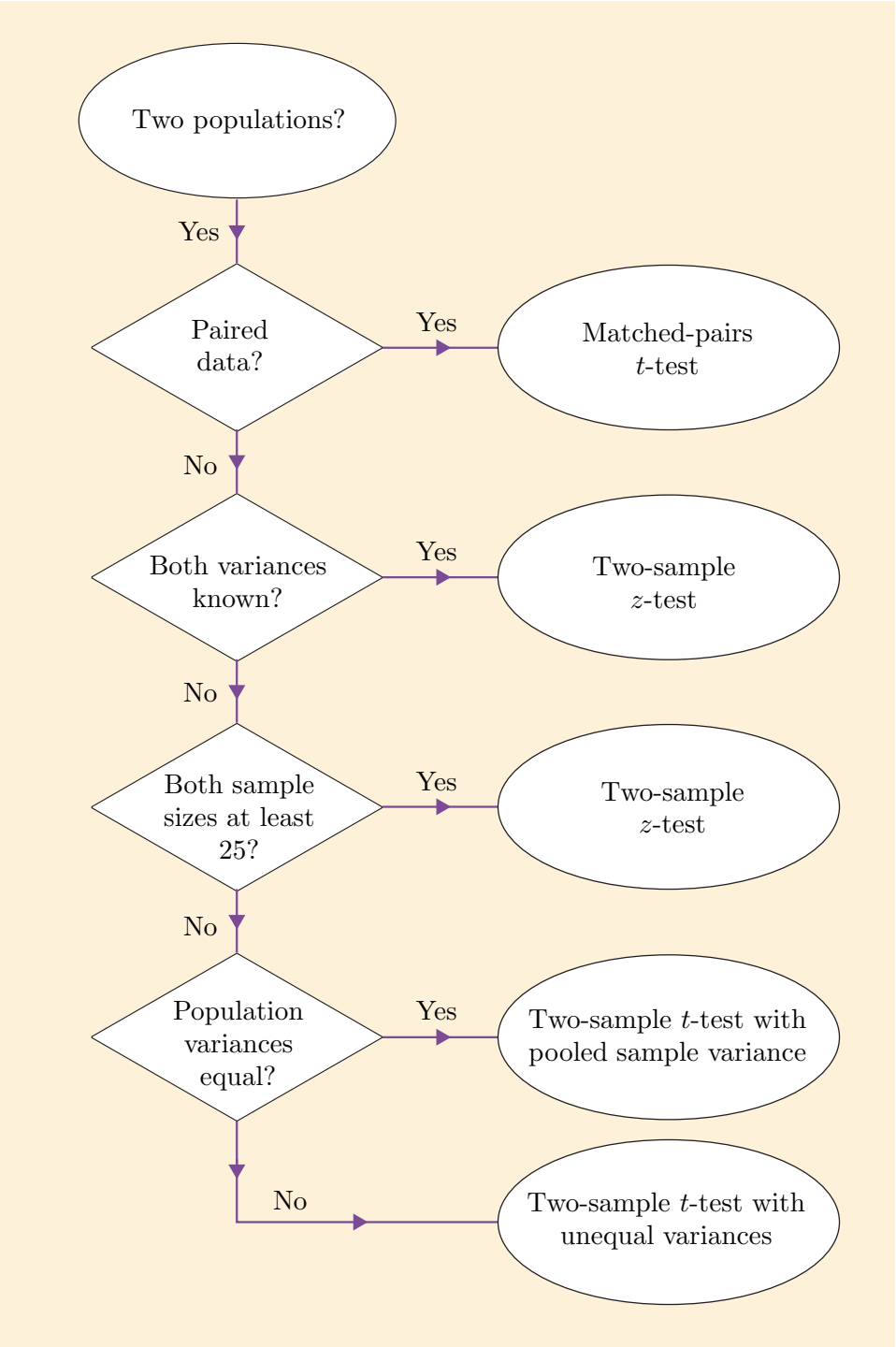


Figure 10 Flow chart for choosing a hypothesis test for inference about the difference between two population means. (Assumptions required for the selected test must also be satisfied.)

Activity 23 Choosing a hypothesis test to compare two means

Samples are taken from two populations in order to test the hypothesis that the population means are equal. What is the appropriate test for each of the following situations? (Assume population distributions are approximately normal, where necessary.) The sample sizes are n_1 and n_2, the population variances are σ_1^2 and σ_2^2, and the sample variances are s_1^2 and s_2^2.

(a) $n_1 = 30$, $n_2 = 50$; σ_1^2 and σ_2^2 are unknown; the data are not matched pairs;

$s_1^2 = 12.1$ and $s_2^2 = 8.5$.

(b) $n_1 = 30$, $n_2 = 14$; σ_1^2 and σ_2^2 are unknown; the data are not matched pairs; $s_1^2 = 8.4$ and $s_2^2 = 9.6$.

(c) $n_1 = 12$, $n_2 = 12$; σ_1^2 and σ_2^2 are unknown; the data are matched pairs; $s_1^2 = 4.8$ and $s_2^2 = 7.3$.

(d) $n_1 = 10$, $n_2 = 10$; σ_1^2 and σ_2^2 are unknown; the data are not matched pairs; $s_1^2 = 12.1$ and $s_2^2 = 3.5$.

(e) $n_1 = 10$, $n_2 = 12$; σ_1^2 and σ_2^2 are known; the data are not matched pairs; $s_1^2 = 6.5$ and $s_2^2 = 5.9$.

You have now covered the material related to Screencast 3 for Unit 12 (see the M140 website).

Performing the hypothesis tests

After the appropriate test has been selected, the null and alternative hypotheses are specified. The (two-sided) hypotheses for all the tests are given in Table 8. For a one-sample test, the hypothesised value of the population mean (μ) is A. When there are two populations, μ_A and μ_B are the population means and $\mu_d = \mu_A - \mu_B$.

Table 8 Null and alternative hypotheses for the (two-sided) z- and t-tests

	H_0	H_1
One-sample tests	$\mu = A$	$\mu \neq A$
Matched-pairs tests	$\mu_d = 0$	$\mu_d \neq 0$
Other two-sample tests	$\mu_A - \mu_B = 0$	$\mu_A - \mu_B \neq 0$

After specifying H_0 and H_1, the test statistic must be calculated. Table 9 gives this statistic for all the tests.

- If there is only one sample, then the sample size, the sample mean and the standard deviation are n, $\overline{x}$ and s.
- If there are two samples forming matched pairs, then the sample mean and the standard deviation of the differences within a pair are $\overline{d}$ and s and the number of pairs is n.
- If there are two (unmatched) samples, then n_A and n_B denote the sample sizes, $\overline{x}_A$ and $\overline{x}_B$ are the sample means, and s_A and s_B are the sample standard deviations.

  When two population variances are assumed to be equal, the pooled estimate of their common standard deviation is

$$s_{\mathrm{p}} = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}}.$$

If the population standard deviation or population standard deviations are known, they will replace the corresponding sample values (this can only apply when the test statistic equals z).

Table 9 Estimated standard error (ESE) and test statistic for the z- and t-tests

Test	ESE	Test statistic
1. One-sample z-test	$\dfrac{s}{\sqrt{n}}$	$z = \dfrac{\overline{x} - A}{\text{ESE}}$
2. One-sample t-test	$\dfrac{s}{\sqrt{n}}$	$t = \dfrac{\overline{x} - A}{\text{ESE}}$
3. Two-sample z-test	$\sqrt{\dfrac{s_A^2}{n_A} + \dfrac{s_B^2}{n_B}}$	$z = \dfrac{\overline{x}_A - \overline{x}_B}{\text{ESE}}$
4. Matched-pairs t-test	$\dfrac{s}{\sqrt{n}}$	$t = \dfrac{\overline{d}}{\text{ESE}}$
5. Two-sample t-test with a common variance	$s_{\text{p}}^2\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}$	$t = \dfrac{\overline{x}_A - \overline{x}_B}{\text{ESE}}$
6. Two-sample t-test with unequal variances	$\sqrt{\dfrac{s_A^2}{n_A} + \dfrac{s_B^2}{n_B}}$	$t = \dfrac{\overline{x}_A - \overline{x}_B}{\text{ESE}}$

To examine the strength of evidence against H_0 that the sample data provide, the test statistic is compared to critical values.

- For tests 1 and 3, the test statistic follows a standard normal distribution, assuming H_0 holds. So the critical values for the two-sided test are 1.96 and -1.96 at the 5% significance level, and 2.58 and -2.58 at the 1% significance level. Hence, for example, the null hypothesis is rejected at the 5% significance level if the test statistic is greater than 1.96 or less than -1.96.
- For tests 2, 4 and 5, the test statistic follows a t distribution if H_0 holds. Critical values for the two-sided test at the 5% significance level are given in Table 2 in Subsection 3.3 of Unit 10 (and the Handbook). The number of degrees of freedom equals $n - 1$ for tests 2 and 4, and $n_A + n_B - 2$ for test 5. If the magnitude of the test statistic is greater than the critical value, then the null hypothesis is rejected at the 5% significance level.
- For test 6, the test statistic approximately follows a t distribution if H_0 holds. However, its degrees of freedom is given by a relatively complicated expression, outside the scope of M140.

The result of the hypothesis test should be stated clearly and conclusions drawn that reflect the setting from which the data came. It is also good practice to state any assumptions that have been made. These will involve the randomness and independence of observations and, for samples of modest size, the assumption that variation in a population is adequately modelled by a normal distribution.

Test 6 is the 'new' hypothesis test that is appropriate when population variances appear to be unequal and the sample sizes are not both large. In activities in the Computer Book you will use Minitab to find p-values for this test.

Example 4 Brinell hardness measurements

In a Brinell hardness test, a hardened steel ball is pressed into the material being tested under a standard load. The diameter of the spherical indentation is then measured. A company is about to replace its current steel ball, Ball A, with a

new steel ball, Ball B. Before doing so, it compares the measurements given by the two balls on eight pieces of material. Each piece of material was tested twice, once with each ball, giving the measurements in Table 10.

The company want to examine whether either ball gives, on average, higher measurements than the other.

Table 10 Diameter measurements from Brinell hardness tests

Sample	1	2	3	4	5	6	7	8
Ball A	63	44	62	51	32	53	46	64
Ball B	42	49	48	29	51	45	52	41

A Brinell hardness testing machine

As a pair of measurements were made on each sample, the data are paired. Thus a paired t-test is appropriate (see the flow chart in Figure 10). We put $\mu_d = \mu_A - \mu_B$ where μ_A and μ_B are the population means for the two balls. The hypotheses are:

$$H_0\colon \mu_d = 0 \quad \text{and} \quad H_1\colon \mu_d \neq 0.$$

To obtain the test statistic, the difference (d) between A's reading and B's reading are determined for each sample.

Sample	1	2	3	4	5	6	7	8
Difference (d)	21	-5	14	22	-19	8	-6	23

Then $n = 8$,

$$\sum d = 21 - 5 + 14 + 22 - 19 + 8 - 6 + 23 = 58$$

and

$$\sum d^2 = 21^2 + (-5)^2 + 14^2 + 22^2 + (-19)^2 + 8^2 + (-6)^2 + 23^2 = 2136.$$

Thus

$$\overline{d} = \frac{58}{8} = 7.25$$

and

$$s^2 = \frac{1}{n-1}\left(\sum d^2 - \frac{(\sum d)^2}{n}\right) = \frac{1}{7}\left(2136 - \frac{58^2}{8}\right) \simeq 245.07.$$

These summary statistics give $s \simeq \sqrt{245.07} \simeq 15.655$,

$$\text{ESE} = \frac{s}{\sqrt{n}} \simeq \frac{15.655}{\sqrt{8}} \simeq 5.5349$$

and

$$t = \frac{\overline{d}}{\text{ESE}} \simeq \frac{7.25}{5.5349} \simeq 1.310.$$

The critical value for a t-test is obtained from Table 2 in Subsection 3.3 of Unit 10 (and the Handbook). The number of degrees of freedom is $n - 1 = 7$, so the critical value is 2.365. The value of t is 1.310, which is less than 2.365. Thus H_0 is not rejected at the 5% significance level.

The conclusion is that there is little evidence of one ball giving higher measurements, on average, than the other. This does not mean that the balls

definitely do not differ systematically – a larger experiment might find evidence of a difference between them.

The assumptions underlying the test are that observations are random, each pair of measurements is independent of all other pairs, and the differences (d) are approximately normally distributed.

Activity 24 Cholesterol reduction

In a nutrition experiment, the effectiveness of two high-fibre diets at reducing serum cholesterol levels was examined. Fifty-seven men with high serum cholesterol were randomly allocated to receive an 'oat' diet or a 'bean' diet for 21 days. Table 11 summarises the fall in serum cholesterol levels (before diet − after diet). Test whether there is a difference between the diets in their effects on cholesterol levels. (The data are artificial.)

No good for this experiment: oats and beans together in these cookies

Table 11 Summary statistics for the fall in cholesterol (mg/dl) on two diets

	Sample size	Sample mean	Sample standard deviation
Oat	29	58.3	19.2
Bean	28	46.4	16.5

You have now covered the material related to Screencast 4 for Unit 12 (see the M140 website).

Constructing confidence intervals

The task of forming confidence intervals for a population mean was first addressed in Unit 9. Given a set of data, there are a range of likely values that the population mean might equal. A confidence interval gives a precisely defined range through consideration of hypothesis tests.

Let μ be the population and consider the hypotheses,

$$H_0\colon \mu = A \quad \text{and} \quad H_1\colon \mu \neq A.$$

Then H_0 will be rejected at the 5% significance level for some values of A but not for others.

Confidence intervals

A 95% confidence interval for μ includes all values of A for which we cannot reject H_0 at the 5% significance level.

A 99% confidence interval for μ includes all values of A for which we cannot reject H_0 at the 1% significance level.

Thus a confidence interval contains the plausible values that μ might take. If you took a very large number of samples from a population, then from each sample you could calculate a 95% confidence interval for μ. Most of these intervals would contain the true value of μ, but some would not. The definition of a confidence intervals enables the following precise statements to be made.

About 95% of the confidence intervals will contain the population mean. For the remaining 5%, about 2.5% will give intervals that are completely below the population mean and about 2.5% will give intervals completely above it.

So if you say that a 95% confidence interval includes the population mean, you will be right 95% of the time; you are 95% confident that your statement is correct.

At the same time, *after calculating a confidence interval*, it is wrong to say 'the probability is 0.95 that this confidence interval contains μ.' Once the confidence interval has been calculated, either it contains the value of μ or it does not. In the former case, the probability is 1 that the interval contains μ, while in the latter case, the probability is 0. (If we take lots of samples from a population, the confidence interval will keep changing but the population mean remains the same. Once the confidence interval has been determined, there is nothing left that is random. Therefore, it is only before gathering data that the probability is 0.95 that the *future* 95% confidence interval will contain μ.)

To show how a confidence interval can be obtained from a hypothesis test, consider the one-sample t-test of H_0: $\mu = A$ versus H_1: $\mu \neq A$. From Table 9, the test statistic is

$$t = \frac{\overline{x} - A}{\text{ESE}}.$$

If t_c is the critical value for the 5% significance level, then the hypothesis that A is the population mean is *not* rejected at the 5% level if

$$\frac{\overline{x} - A}{\text{ESE}} \leq t_c \quad \text{and} \quad -t_c \leq \frac{\overline{x} - A}{\text{ESE}}.$$

Thus, it is not rejected at the 5% level if

$$\overline{x} - A \leq t_c \times \text{ESE} \quad \text{and} \quad -t_c \times \text{ESE} \leq \overline{x} - A,$$

which is equivalent to

$$\overline{x} - t_c \times \text{ESE} \leq A \quad \text{and} \quad A \leq \overline{x} + t_c \times \text{ESE}.$$

Thus A is not rejected at the 5% significance level if it is in the interval

$$(\overline{x} - t_c \times \text{ESE},\ \overline{x} + t_c \times \text{ESE}).$$

By definition, this interval is the 95% confidence interval for the population mean. It is an example of the following more general result.

Confidence interval for a mean or the difference between two means

The lower limit of the confidence interval is:

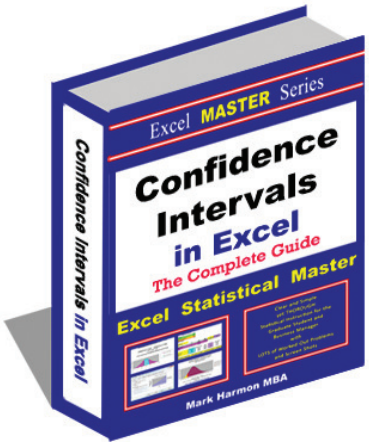
point estimate $-$ (z or t critical value) $\times$ ESE

and the upper limit is:

point estimate $+$ (z or t critical value) $\times$ ESE,

where ESE is the estimated standard error of the point estimate.

To apply this result requires a point estimate, a z or t critical value, and an ESE. The point estimate will be the sample mean when making inferences about one



Confidence intervals: they're everywhere!

population mean, and it will be the difference between the two samples means when making inferences about the difference between two population means. Table 9 gives the ESE. Actually, it gives lots of ESEs. To decide which one is appropriate:

- If the interval is for a population mean, consider what hypothesis test you would use to test H_0: $\mu = A$. The ESE for that hypothesis test is the one that should be used to form the confidence interval.
- If the interval is for the difference between two population means, consider what hypothesis test you would use to test H_0: $\mu_A = \mu_B$. The ESE for that hypothesis test is the one that should be used to form the confidence interval.

The choice of hypothesis test also determines the z or t critical value that is used to form the confidence interval. If a z-test is the appropriate hypothesis test, then the z-value of 1.96 should be used for a 95% confidence interval and 2.58 for a 99% confidence interval. If a t-test is the appropriate hypothesis test, then t_c should be obtained from Table 2 in Subsection 3.3 of Unit 10 (and the Handbook); the degrees of freedom (as with the hypothesis tests) are $n - 1$ for the one-sample t-test and paired t-test, and $n_A + n_B - 2$ for the two-sample t-test with pooled sample variance.

Example 5 Confidence interval from Brinell hardness measurements

Example 4 concerned measurements from Brinell hardness tests using two hardened steel balls, Ball A and Ball B. A paired t-test was used to test whether the difference between the population means (μ_A and μ_B) was 0.

Suppose, now, that a 95% confidence interval for $\mu_A - \mu_B$ is required. Using the data in Example 4, the mean for Ball A is

$$\frac{63 + 44 + 62 + 51 + 32 + 53 + 46 + 64}{8} = 51.875$$

and the mean for Ball B is

$$\frac{42 + 49 + 48 + 29 + 51 + 45 + 52 + 41}{8} = 44.625.$$

Hence the point estimate is $51.875 - 44.625 = 7.25$. (This equals $\overline{d}$, of course, which was calculated in Example 4. The separate means for Ball A and Ball B did not have to be calculated and $\overline{d}$ could have been taken as the point estimate.)

From Example 4, ESE $\simeq 5.5349$. The degrees of freedom are $n - 1 = 7$ and, for 7 degrees of freedom, 2.365 is the critical value, t_c. The value of 2.365 could be obtained from Table 2 in Subsection 3.3 of Unit 10 (and the Handbook), but it has already been obtained in Example 4.

Hence the lower limit of the 95% confidence interval is

$$7.25 - 2.365 \times 5.5349 \simeq -5.8$$

and the upper limit is

$$7.25 + 2.365 \times 5.5349 \simeq 20.3,$$

so the 95% confidence interval for $\mu_A - \mu_B$ is $(-5.8, 20.3)$.

In the last example, notice that the 95% confidence interval contains the value 0 – so the hypothesis that the population means are equal would not be rejected at the 5% significance level. This was also the conclusion from the hypothesis test

in Example 4, and it illustrates the following close connection between confidence intervals and hypothesis tests.

- If the 95% confidence interval does not include the value given by the null hypothesis, we reject the null hypothesis at the 5% significance level.
- If the 99% confidence interval does not include the value given by the null hypothesis, we reject the null hypothesis at the 1% significance level.

Activity 25 Confidence intervals for cholesterol reduction

Data from a nutrition experiment were summarised in Activity 24. The data related to the reduction in serum cholesterol level on two diets, an 'oat' diet (diet A) and a 'bean' diet (diet B). Let μ_A and μ_B denote the population mean reductions on the two diets.

(a) Construct a 95% confidence interval for $\mu_A - \mu_B$.

(b) Construct a 99% confidence interval for $\mu_A - \mu_B$.

(c) Which interval is shorter? Would that always be the shorter interval?

(d) Which of the confidence intervals contain the value 0? How does that relate to the results of the hypothesis test in Activity 24?

You have now covered the material needed for Subsection 12.2 of the Computer Book.

Exercises on Section 5

Exercise 8 Which one-sample test?

A sample is taken from a population in order to test the hypothesis that the population mean equals 50. What is the appropriate test for each of the following situations, assuming the population distribution is approximately normal?

(a) The sample size is 20 and the population variance is known.

(b) The sample size is 12 and the population variance is unknown.

(c) The sample size is 40 and the population variance is unknown.

(d) The sample size is 35 and the population variance is known.

Exercise 9 Which test for comparing two means?

Samples are taken from two populations in order to test the hypothesis that the population means are equal. What is the appropriate test for each of the following situations? (Assume population distributions are approximately normal, where necessary.) The sample sizes are n_1 and n_2, the population variances are σ_1^2 and σ_2^2, and the sample variances are s_1^2 and s_2^2.

(a) $n_1 = 8$, $n_2 = 18$; σ_1^2 and σ_2^2 are unknown; the data are not matched pairs; $s_1^2 = 23.8$ and $s_2^2 = 28.2$.

(b) $n_1 = 20$, $n_2 = 20$; σ_1^2 and σ_2^2 are unknown; the data are matched pairs; $s_1^2 = 1.3$ and $s_2^2 = 1.7$.

(c) $n_1 = 50$, $n_2 = 9$; σ_1^2 and σ_2^2 are unknown; the data are not matched pairs; $s_1^2 = 1.7$ and $s_2^2 = 9.1$.

(d) $n_1 = 30$, $n_2 = 30$; σ_1^2 and σ_2^2 are unknown; the data are not matched pairs; $s_1^2 = 11.3$ and $s_2^2 = 15.5$.

(e) $n_1 = 10$, $n_2 = 30$; σ_1^2 and σ_2^2 are unknown; the data are not matched pairs; $s_1^2 = 17.4$ and $s_2^2 = 10.6$.

6 Correlation and regression

Often, the reason for gathering data is to learn about the relationship between different variables. When only two variables are involved, much can be learned about their relationship by drawing a scatterplot of one variable against the other. In Subsection 6.1, we consider the information that a scatterplot can provide. In Subsection 6.2, correlation and regression are reviewed.

6.1 Scatterplots and relationships

Data are said to be linked when two or more variables are recorded for the same sampling units. Here, the focus is on the case where there are two variables, so that the linked data are paired data.

Scatterplots are a useful tool for examining the relationship between a pair of variables. They can address the following question/s.

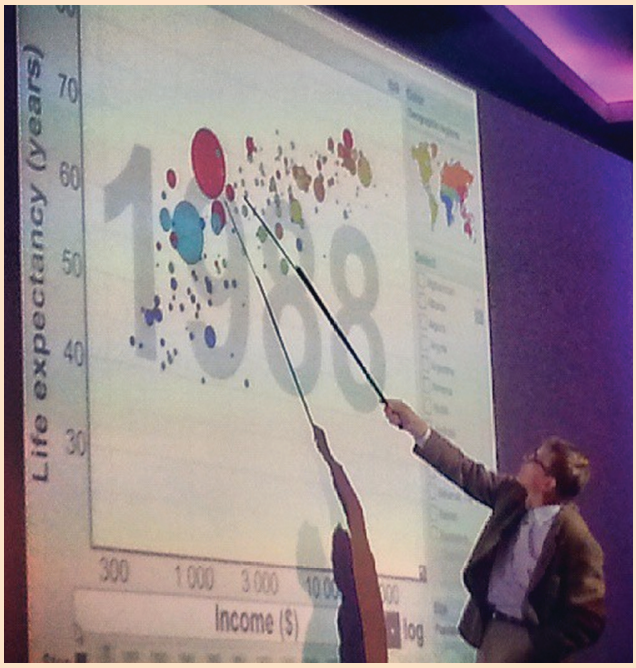
Is there a relationship between the two variables? If so:

- Is the relationship positive or negative, or is it neither?
- Is the relationship strong or weak?
- Is the relationship linear?

Professor Hans Rosling (b. 1948)

Hans Rosling is a Swedish public health doctor, academic and statistician. He began his career in public health, spending some time in remote rural parts of Africa. Since 1997, he has been Professor of International Health at the Karolinska Institute, a world-renowned medical university in Stockholm.

More recently, Rosling has become famous for persuasively presenting data and ideas on health, international development and many other things. With his son and daughter-in-law he set up the Gapminder Foundation in 2005, to develop animated software for showing data and to use it to show global development trends.



Professor Hans Rosling does great things with scatterplots!

Figure 11 plots, for ten towns, the percentage of households in owner-occupied housing against the percentage of employed residents working in manufacturing.

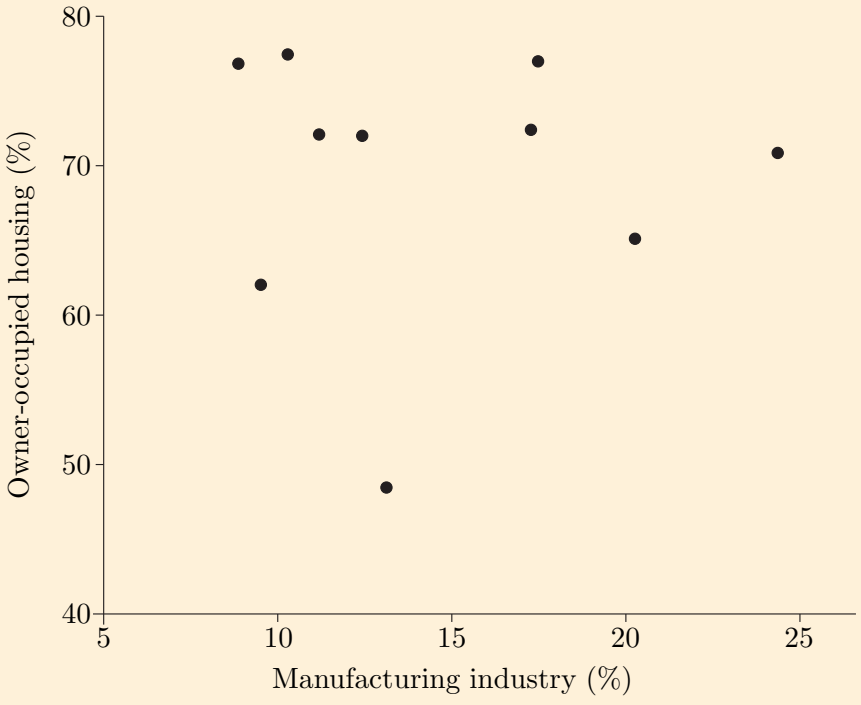


Figure 11 Percentage of employed residents working in manufacturing and percentage of households in owner-occupied houses

There appears to be no relationship between the variables: if in an eleventh town you knew the percentage of employed residents working in manufacturing, it would not give any indication of the likely percentage of households in that town who live in owner-occupied housing. Similarly, knowing the percentage of

households in owner-occupied housing would give no indication of the percentage of employed residents who work in manufacturing.

There is said to be no relationship between two variables when knowledge of one of them provides no information about the value of the other.

Figure 12 plots two variables that are related. The shaded area contains the points and it slopes upwards: as the percentage of unemployed men in a town increases, the percentage of households with no car also increases. Thus, if in an eleventh town the percentage of men unemployed were known, it would influence an estimate of the likely percentage of households in the town that had no car.

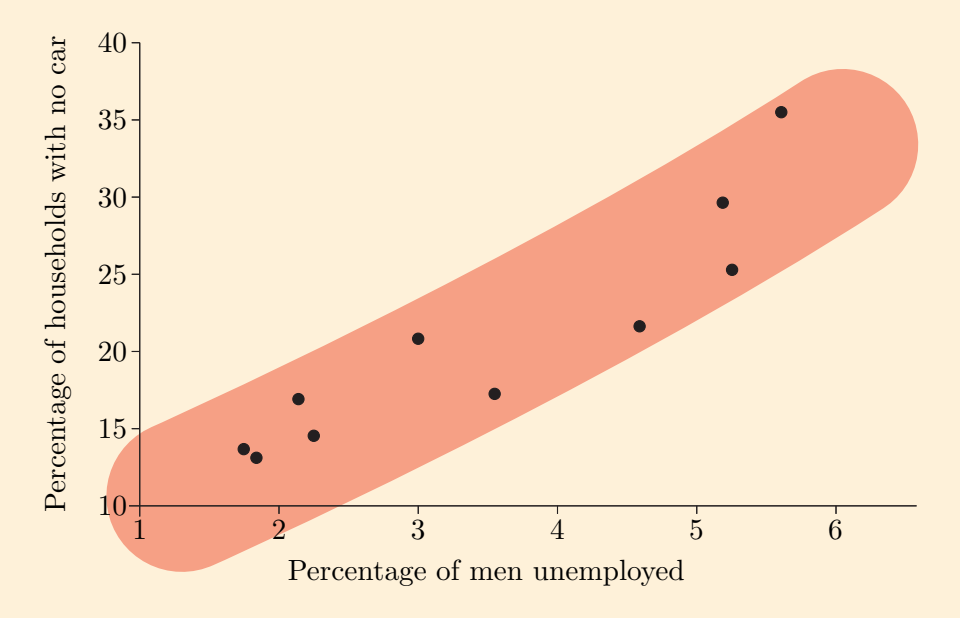


Figure 12 Percentage of males unemployed and percentage of households with no car in ten towns

The variables in Figure 12 are said to be *positively related* because an increase in one variable is associated with an increase in the other variable. Variables are said to be *negatively related* if an increase in one variable is associated with a decrease in the other variable. An example of negatively related variables is given in Figure 13.

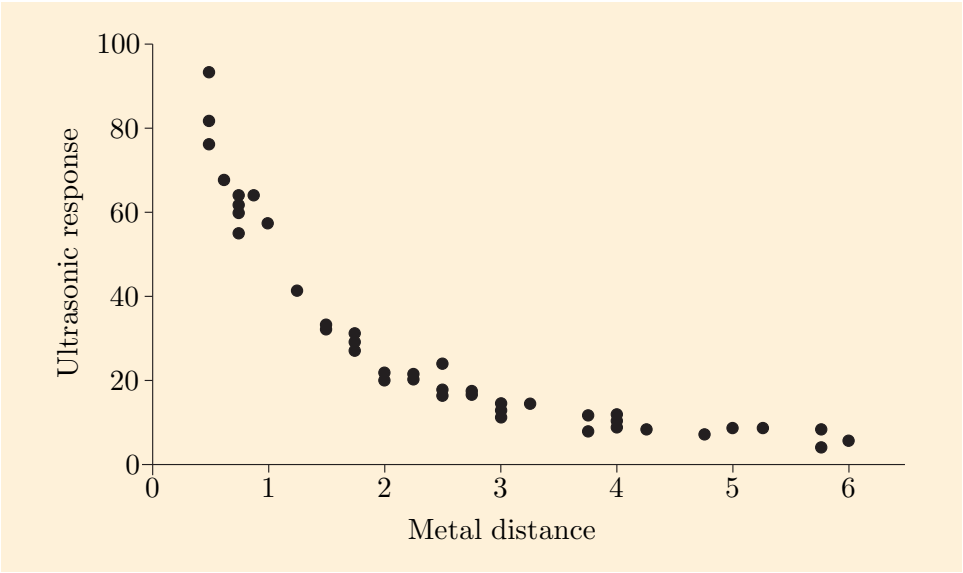


Figure 13 Data from an ultrasonic calibration study

Variables need not be positively or negatively related, even when there is clearly a relationship between them. This is illustrated in Figure 14, where y is large for values of x near 10 and values of x near 20, and smaller for values of x near 15.

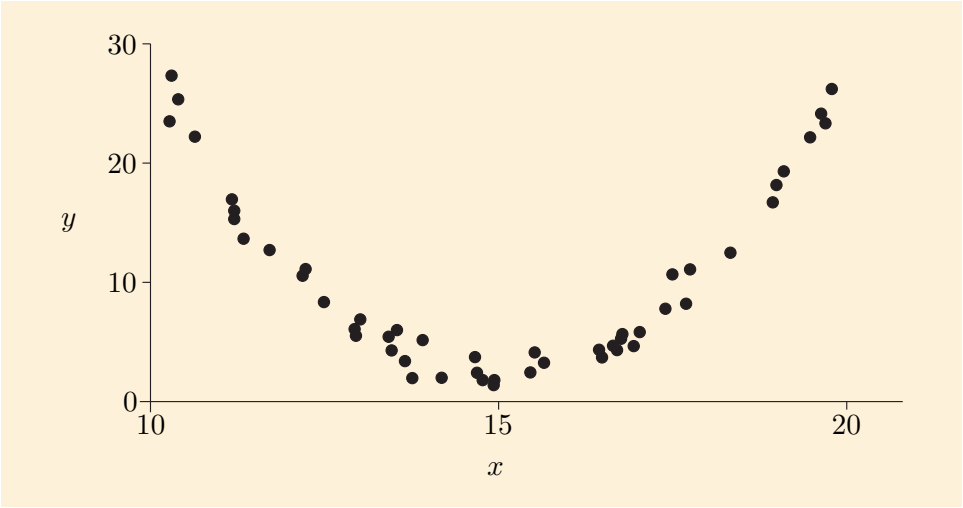


Figure 14 A scatterplot of some data

A relationship is said to be *strong* when all the points on a scatterplot lie close to a line.

A relationship is said to be *weak* when all the points only loosely follow a line.

The relationships in Figures 12, 13 and 14 are all strong. Figure 15 is an example of a weak (positive) relationship.

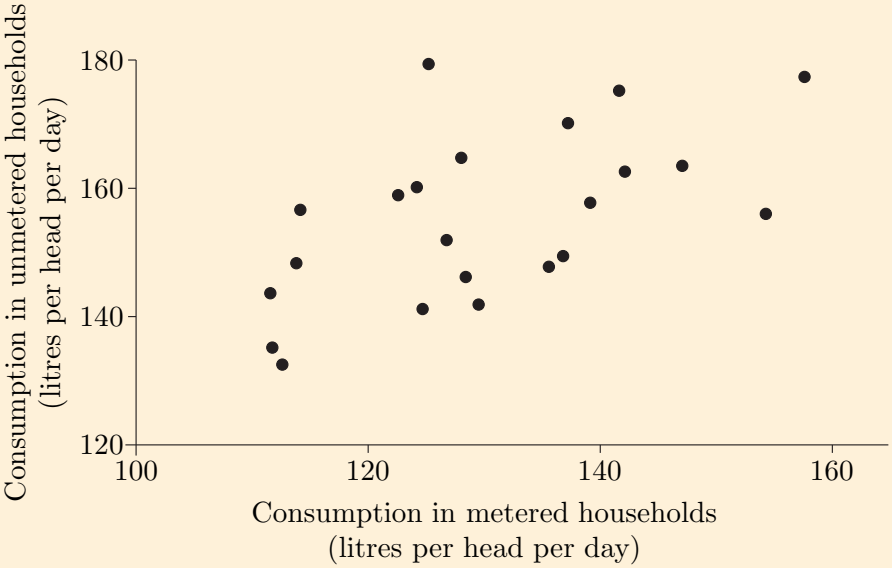
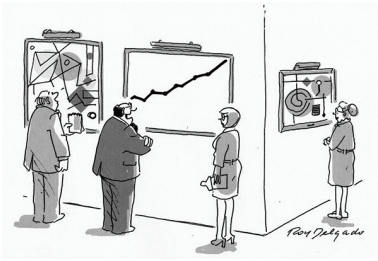


Figure 15 Average water consumption in metered and unmetered households

An important characteristic of a relationship is whether it is linear or non-linear. A relationship is said to be linear if it can be summarised reasonably well by a straight line. It is said to be non-linear if it can be summarised reasonably well by a curve but not by a straight line. Consequently, the relationships in Figures 13 and Figure 14 are non-linear, while Figure 12 shows a linear relationship. This is not typical though – linear relationships commonly occur in practice.


'I'll take it.'

Activity 26 Identifying characteristics of relationships

For each of the following six scatterplots, say whether there is a relationship between the variables. If there is one, classify it as (i) positive, negative or neither, (ii) weak or strong, and (iii) linear or non-linear.

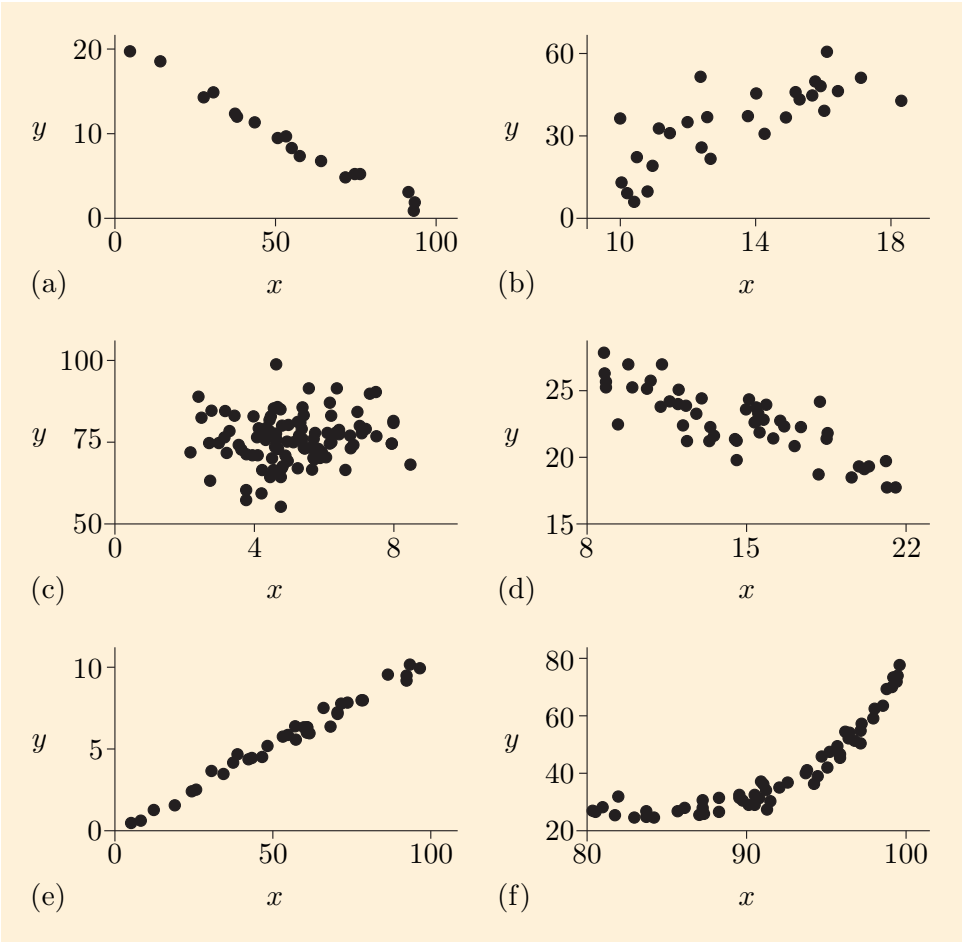


Figure 16 Six scatterplots showing different relationships

6.2 Linear relationships

Correlation *measures* the strength of a linear relationship.

Regression *describes* a linear relationship.

Correlation

The correlation coefficient has the following formula:

$$\text{Correlation} = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \times \sum(y-\overline{y})^2}}.$$

The value of the correlation coefficient is always between -1 and $+1$. A value near $+1$ means there is a strong positive relationship between x and y and a value near -1 indicates a strong negative relationship between them. For example, the variables in Figure 12 (Subsection 6.1) show quite a strong positive relationship. Their correlation coefficient (calculated using the above formula and the data that gave the figure) is 0.91, which is reasonably close to 1. In contrast, the variables in Figure 15 have a relationship that is weak (and positive) with the correlation coefficient lower at $\simeq 0.57$. When there is little relationship between two variables, the correlation coefficient takes a value near 0. Thus in Figure 11, the two variables appear unrelated and the correlation coefficient is -0.01.

The correlation coefficient only reflects the degree to which a relationship is linear. In Figure 14, there is a very strong relationship but the relationship is non-linear – the correlation coefficient is 0.000 (correct to three decimal places). However, with many non-linear relationships, the relationship can *partly* be modelled by a straight line. For example, in Figure 13, the relationship is clearly non-linear, but a straight line with a negative slope would partly capture the negative relationship between the two variables. The correlation coefficient is -0.85, which is far from 0.

Activity 27 Ordering correlation coefficients.

Order the six scatterplots in Figure 16 (Subsection 6.1) according to the correlation between the variables in the plots. The scatterplot corresponding to the highest correlation coefficient should be first in the list, the one corresponding to the second highest correlation coefficient should be second in the list, and so on. Thus the one corresponding to the most negative correlation coefficient should be at the end of the list.

The following is the procedure for calculating the correlation coefficient.

Calculating the correlation coefficient

Given a batch of n linked data pairs, (x, y), the correlation coefficient (r) is obtained as follows:

1. Calculate $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$ and $\sum xy$.
2. Calculate
$$\sum (x-\overline{x})^2 = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2,$$
$$\sum (y-\overline{y})^2 = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2,$$
$$\sum (x-\overline{x})(y-\overline{y}) = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right).$$
3. Use the values from step 2 to calculate
$$r = \frac{\sum (x-\overline{x})(y-\overline{y})}{\sqrt{\sum (x-\overline{x})^2 \times \sum (y-\overline{y})^2}}.$$

Example 6 Basal area and weight of trees

Sitka spruce

During crop-thinning of a small forest of Sitka spruce, seven trees were felled and the cross-sectional area of each tree at its base (its *basal area*), x, and total dry weight, y, were measured. The following results were obtained.

Table 12 Basal area and dry weight of seven trees

x (m^2 $\times$ 100)	2.24	1.06	0.79	1.78	1.22	0.54	1.40
y (kg)	79.1	33.9	19.2	54.8	51.0	12.0	46.7

To determine the correlation coefficient between x and y we first calculate:

$$\sum x = 2.24 + 1.06 + \ldots + 1.40 = 9.03$$

$$\sum y = 79.1 + 33.9 + \ldots + 46.7 = 296.7$$

$$\sum x^2 = 2.24^2 + 1.06^2 + \ldots + 1.40^2 = 13.6737$$

$$\sum y^2 = 79.1^2 + 33.9^2 + \ldots + 46.7^2 = 15\,703.59$$

$$\sum xy = 2.24 \times 79.1 + 1.06 \times 33.9 + \ldots + 1.40 \times 46.7 = 459.91.$$

The sample size is $n = 7$, so

$$\sum (x - \overline{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$
$$= 13.6737 - \frac{(9.03)^2}{7} = 2.025.$$

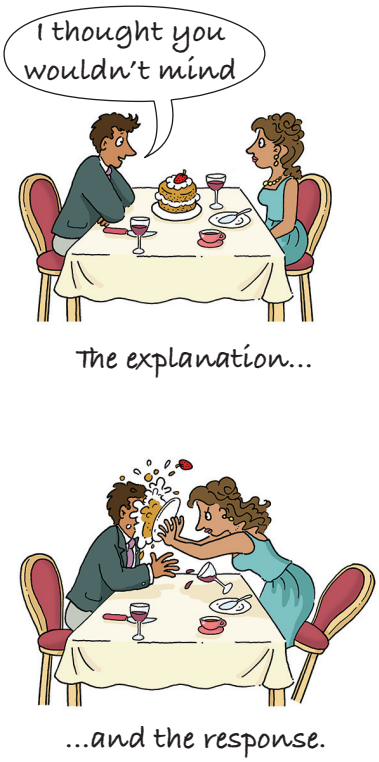
$$\sum (y - \overline{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$
$$= 15\,703.59 - \frac{(296.7)^2}{7} \simeq 3127.75.$$

$$\sum (x - \overline{x})(y - \overline{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$
$$= 459.91 - \frac{9.03 \times 296.7}{7} = 77.167.$$

Hence in this example,

$$\text{correlation} = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sqrt{\sum (x - \overline{x})^2 \times \sum (y - \overline{y})^2}}$$
$$\simeq \frac{77.167}{\sqrt{2.025 \times 3127.75}} \simeq 0.97.$$

This correlation is close to 1, indicating a strong linear relationship between basal area and dry weight.



Regression

Data on Sitka spruce trees were given in Example 6. The dataset could be used to obtain an equation for estimating a tree's dry weight from the cross-sectional area at its base. The equation is potentially useful since, while a tree is growing, its basal area is much easier to measure than its weight. A suitable equation can be obtained through least squares linear regression, provided the relationship between the variables is linear.

In considering the correlation between two variables, it does not matter which variable is called x and which is called y – swapping them around would not change the value of the correlation coefficient. That is, in correlation the two variables have identical roles. In regression though, there is a response variable and an explanatory variable, and these have different roles.

A *response* variable (usually denoted as y) is the variable that is being explained or whose value depends on the other variable. It is also the variable to be predicted if predictions are to be made.

An *explanatory* variable (usually x) is the variable that is doing the explaining or is the variable on which the response variable depends.

Suppose now that the dry weight of a Sitka spruce is to be estimated from its basal area. Then the dry weight is y and the basal area is x. In Example 7, we will obtain the least squares regression line relating basal area to dry weight. The

line is

$$y = -6.77 + 38.1x.$$

Given a tree's basal value, this equation gives its fitted value:

$$\text{dry weight fitted value} = -6.77 + 38.1 \times \text{basal area}.$$

For example, the first tree in the dataset has a basal area of 2.24, so its fitted value is

$$-6.77 + 38.1 \times 2.24 \simeq 78.6.$$

If we did not know the tree's actual dry weight, then it would be estimated as 78.6.

The following table extends Table 12, from Example 6, to include the fitted values for the seven trees in the dataset.

Table 13 Basal areas, actual dry weights and fitted values of seven Sitka spruce

Basal area (x)	2.24	1.06	0.79	1.78	1.22	0.54	1.40
Actual dry weight (y)	79.1	33.9	19.2	54.8	51.0	12.0	46.7
Fitted value	78.6	33.6	23.3	61.0	39.7	13.8	46.6

Figure 17 is a scatterplot of these data and also shows the least squares regression line. The actual values of the data are marked by black dots. It can be seen that the regression line virtually passes through the third, fifth and seventh data points, and is close to the others. Hence the line fits the data well. (In Example 6, the correlation coefficient of 0.97 indicated a strong linear relationship between basal area and dry weight, so it could be anticipated that the line would fit the data well.)

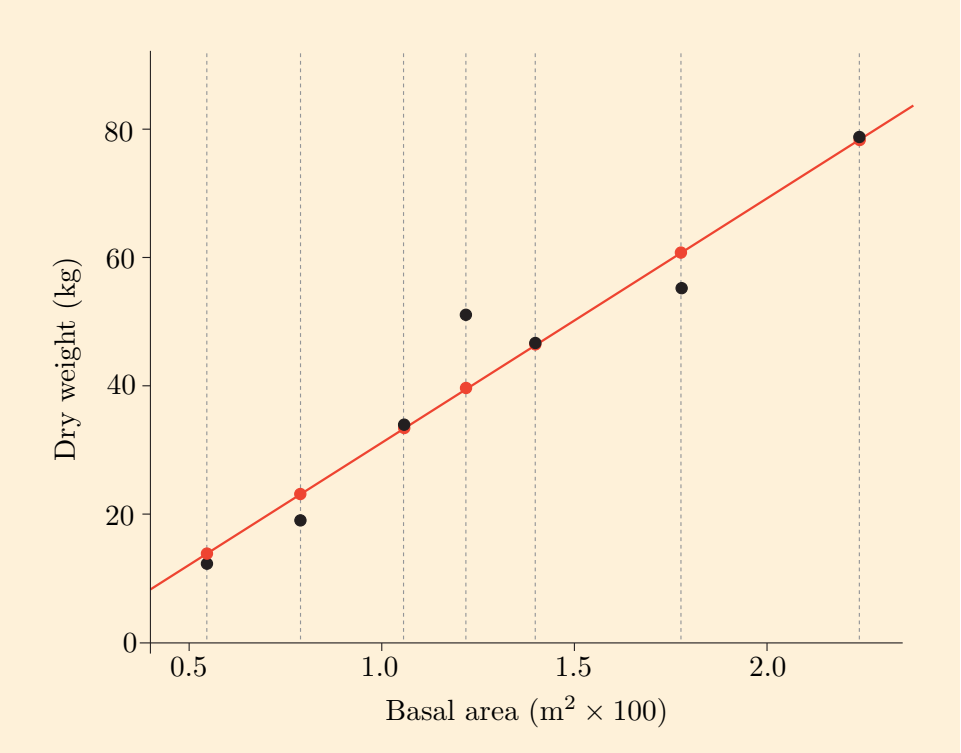


Figure 17 Plot of dry weight against basal area, with least squares regression line

Residuals are defined as the differences between the fitted values and the data values:

$$\text{Residual} = \text{Data} - \text{Fit}.$$

In Figure 17, the fitted values of the seven trees are shown as orange dots. From Table 13, their positions are $(2.24, 78.6), (1.06, 33.6), \ldots, (1.40, 46.6)$. As the actual data values (the black dots) are at $(2.24, 79.1)$, $(1.06, 33.9), \ldots, (1.40, 46.7)$, the data points lie vertically above the fitted values, or vertically below them. Thus, residuals are the vertical distances from the data points to the line.

A regression line can serve various purposes, but to understand why it is constructed in the way that it is, think of the line as a way of predicting the response variable, y, when we know the value of the explanatory variable, x. Each residual is the inaccuracy in predicting a y-value in the dataset, given its corresponding x-value. Thus, the residuals indicate the usefulness of the line for making predictions. The aim in least squares regression is to minimise the sum of the squares of these residuals. Thus, in deciding how close a line is to the data points, the *perpendicular* distances from the data points to the line are not the quantities of interest – the focus is on the *vertical* distances from each data point to the line. The least squares regression is calculated as follows.

Calculation of the least squares regression line $y = a + bx$ for a set of n data points (x, y)

1. Calculate $\sum x$, $\sum y$, $\sum (x - \overline{x})^2$ and $\sum (x - \overline{x})(y - \overline{y})$.
2. Calculate the means of x and y:
$$\overline{x} = \frac{\sum x}{n} \quad \text{and} \quad \overline{y} = \frac{\sum y}{n}.$$
3. The slope b is given by
$$b = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sum (x - \overline{x})^2}.$$
4. The intercept a is given by
$$a = \overline{y} - b\overline{x}.$$

For step 1 above, $\sum (x - \overline{x})^2$ and $\sum (x - \overline{x})(y - \overline{y})$ can be obtained from

$$\sum (x - \overline{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

and

$$\sum (x - \overline{x})(y - \overline{y}) = \sum xy - \frac{\sum x \sum y}{n}.$$

However, if you have calculated the correlation coefficient, you will already have obtained these quantities as part of the procedure for its computation. (This reflects the close connection between correlation and regression.)

Example 7 Regression line relating a tree's basal area and weight

To determine the least squares regression line relating a Sitka spruce's basal area to its dry weight, the following quantities obtained in Example 6 will be used.

$$\sum x = 9.03, \quad \sum y = 296.7, \quad n = 7,$$
$$\sum (x - \overline{x})^2 = 2.025, \quad \sum (x - \overline{x})(y - \overline{y}) \simeq 77.167.$$

These give

$$\overline{x} = \frac{\sum x}{n} = \frac{9.03}{7} = 1.29 \quad \text{and} \quad \overline{y} = \frac{\sum y}{n} = \frac{296.7}{7} \simeq 42.386.$$

Hence, the slope is

$$b = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sum (x - \overline{x})^2} = \frac{77.167}{2.025} \simeq 38.107,$$

and the intercept is

$$\begin{aligned} a = \overline{y} - b\overline{x} &= 42.386 - 38.107 \times 1.29 \\ &\simeq 42.386 - 49.158 = -6.772. \end{aligned}$$

These form the regression line:

$$y = -6.77 + 38.1x.$$

A regression line can be used to form estimates and make predictions. Standard terminology is to *estimate* a mean response and *predict* an individual response. The last example will be used to discuss the two cases.

Given a particular basal area, we might want to:

1. Estimate the mean dry weight of Sitka spruce that have that basal area.
2. Predict the dry weight of a single Sitka spruce tree that has that basal area.

In both cases, the estimate/prediction will be the point on the regression line that corresponds to the specified basal value, i.e. $-6.77 + 38.1x$, where x is the specified basal area. However, accuracy will not be the same in the two cases.

To elaborate, if the position of the regression line were known precisely, then the mean dry weight of Sitka spruce trees that have a particular basal area could be estimated with perfect accuracy. But there would still be uncertainty in predicting the dry weight of a single Sitka spruce tree from its basal area, because individual observations do not lie on the regression line – they vary around it.

More generally, if we took another sample of seven Sitka spruce, the regression line would not be identical to the one we have calculated. That is, the position of the regression line is affected by random variation. This random variation affects both the estimate of a mean response and the prediction of a single response. It is the only source of uncertainty that affects the estimated mean response, whereas individual variation also affects the accuracy with which a single response is predicted. Consequently, for any given x, there is greater uncertainty in predicting an individual response than in estimating a mean response.

Given a particular x, a confidence interval for the mean response can be determined using methods similar to those described in Subsection 5.2 (though the ESE is more complicated). Different values of x can be considered and the corresponding confidence interval determined for each. As x is varied, the confidence interval changes smoothly, getting narrower towards the mean of the x-values. This is illustrated in Figure 18, where long-dashed lines map

end-points of the 95% confidence intervals for the mean dry weight as the basal area (x) changes.

An interval estimate for a prediction is referred to as a *prediction interval*. The dotted lines in Figure 18 map end-points of the 95% prediction intervals for an individual response. It can be seen that the prediction intervals are much wider than the confidence intervals – much of the uncertainty in predicting an individual value stems from the random variation of individual values about the regression line.

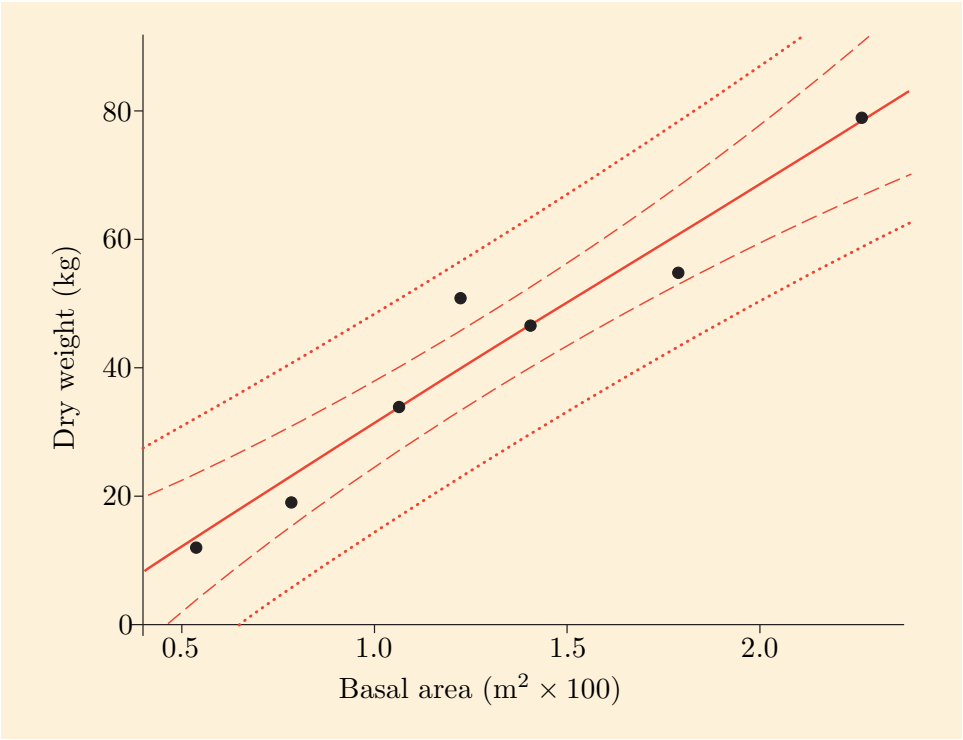


Figure 18 95% confidence intervals and 95% prediction intervals for dry weight as basal area varies

Figure 18 shows that the relationship between basal area and dry weight is approximately linear *when the basal area is between 0.5 and 2.3*. The data do not tell us whether the linear relationship extends outside that range. Hence, the regression line should not be used to make predictions (or estimate the mean response) for basal areas below 0.5 or above 2.3. More generally, making predictions outside the range of x-values in the original sample is termed **extrapolation** and should be avoided, as the validity of the predictions or prediction intervals would be unknown.

Exercises on Section 6

Exercise 10 Checkout time

The time taken (y seconds) to deal with a customer at a supermarket checkout consists of a basic part plus an amount that depends on the number of items bought. Observations of the time taken for each of 30 customers were recorded, together with the number of items (x) that the customer had bought. The following summarises the data:

$$\sum x = 620, \quad \sum y = 2776, \quad \sum xy = 79\,136$$
$$\sum x^2 = 19\,378, \quad \sum y^2 = 344\,742, \quad n = 30.$$

(a) Calculate the correlation coefficient between the time at the checkout and the number of purchases, based on this information.

(b) Does the correlation coefficient suggest that a least squares regression line would be a good way of representing the relationship between time at the checkout and number of purchases? What further information would you need in order to answer this question with more confidence?

Exercise 11 Relating number of purchases to checkout time

Using the data in Exercise 10, estimate the least squares regression line,

$$y = a + bx.$$

Interpret the slope and intercept of this line in the context of the time taken to serve a customer.

7 Computer work: binomial and t-test

In Subsection 3.2, you learned the general form of the binomial distribution. In this section, you will explore the shape of this distribution for different values of n and p. You will also use Minitab to perform an unpaired t-test to compare two population means when it is *not* assumed that the two population variances are equal. You should work through all of Chapter 12 of the Computer Book now, if you have not already done so.

Summary

In this unit, we have reviewed the main themes of M140: numerical and graphical summaries of data, the collection of data, probability, hypothesis testing, confidence intervals, correlation and regression. You may have seen more clearly the similarities and links between many of the concepts and methods that you have learned in the module. You will also have gained more practice at using many of these methods.

You have learned that the probabilities you calculated in Unit 6 for the sign test are probabilities from a binomial distribution in which $p = \frac{1}{2}$. You have learned how to calculate binomial probabilities for other values of p and to identify situations where the binomial distribution arises. You have also used Minitab to calculate binomial probabilities and explored the shape of the binomial distribution for different sample sizes and values of p.

There are a variety of situations in which z-tests and t-tests can be used to test whether a population mean has a particular value or if two population means are equal. You have learned how to choose the appropriate test for the different situations. You have also learned about a two-sample t-test that does not assume population variances are equal. You have learned to use Minitab to perform the test.

Learning outcomes

After working through this unit, you should be able to:

- interpret a growth chart
- combine different sampling methods in designing a survey
- identify situations where a binomial distribution applies
- calculate probabilities for a binomial distribution, both by hand and using Minitab
- choose an appropriate test for making inferences about a population mean
- choose an appropriate test for making inferences about the difference between two population means
- recognise the similarity of the test statistic for different forms of z-test and t-test
- describe the two-sample t-test for comparing two population means when the population variances are unequal and perform the test using Minitab.

Also, this unit has reviewed many learning objectives from the first 11 units of M140 – more than you might have thought. Working though the unit will have consolidated your ability to:

- find the mean and median of a batch of data
- find the upper and lower quartiles and the interquartile range of a batch of data
- calculate the variance and standard deviation of a batch of data
- prepare a five-figure summary of a batch of data
- draw and interpret the boxplot of a batch of data
- draw a stemplot of a batch of data
- use stemplots and boxplots to decide whether a batch of data is symmetric, left-skew or right-skew
- appreciate the priorities in summarising a batch of data
- find the weighted mean of a set of numbers with associated weights
- describe the major steps in producing the Retail Prices Index
- calculate a simple chained price index and explain what is meant by its base date
- describe quota sampling in general terms
- choose a simple random sample using random numbers and a labelled list of the target population
- choose a systematic random sample using random numbers and a labelled list of the target population
- describe the relative strengths and weaknesses of simple and systematic random sampling
- describe the principles involved in cluster sampling and stratified sampling
- choose a random sample for a stratified survey using random numbers and a labelled list of the target population
- choose a random sample for cluster sampling using random numbers and a labelled list of the target population
- distinguish between three kinds of experiments (exploratory, measurement and hypothesis testing)

- appreciate the need to randomise and reduce sources of bias when designing an experiment
- explain why placebos are used in clinical trials and the purpose of double-blind trials
- describe crossover, matched-pairs and group-comparative designs of clinical trials
- calculate probabilities based on random selection
- calculate joint and conditional probabilities
- state and use the relationship between joint and conditional probabilities
- express the independence of two events in terms of conditional probabilities
- state and use the general addition rule for probabilities
- count the number of ways that a specified combination can occur
- understand the concepts of a hypothesis test and the main steps in performing a hypothesis test
- carry out the χ^2 test for contingency tables, taking account of the size of Expected values
- interpret a χ^2 test in terms of the null and alternative hypotheses
- decide when to use the χ^2 test for contingency tables
- appreciate that we can think of all normal distributions in terms of the standard normal distribution
- apply the formula that transforms any variable x with a given normal distribution to the variable z with the standard normal distribution
- appreciate that, whatever the shape of the population distribution, for a large enough sample size the sampling distribution of the mean is nearly always approximately normal
- write down the mean and standard deviation of the sampling distribution of the mean for samples of size n, given the population mean, μ, and standard deviation, σ
- carry out a two-sample z-test to analyse the difference between means
- carry out a matched-pairs t-test
- examine whether it is reasonable to assume that two population variances are equal
- carry out a two-sample t-test when population variances are equal
- appreciate the relationship between hypothesis tests and confidence intervals
- calculate a confidence interval for a population mean
- interpret a confidence interval
- calculate a confidence interval for the difference between two population means
- explain what is meant by a relationship between two variables
- recognise positive and negative relationships from a scatterplot
- describe a relationship between two variables which is neither positive nor negative
- recognise strong and weak relationships from a scatterplot
- understand the concept of the correlation coefficient – in particular, how it relates to a relationship between two variables shown on a scatterplot

- calculate the correlation coefficient by hand
- understand the terms response variable and explanatory variable, and decide which is which in a given example
- calculate a least squares regression line for a batch of linked data by hand
- use a regression line to predict the value of the response variable, and know when it is appropriate to do this
- interpret a confidence interval for estimating the mean response using a least squares regression line
- interpret a prediction interval for individual predictions from a least squares regression line.

Take a bow for reaching the end of the last unit!

The module team

The following team worked on the production of M140 *Introducing statistics*.

Module team chairs

Paul Garthwaite and Karen Vines

Academic contributors

Paddy Farrington, Chris Jones, Kevin McConway, Heather Whitaker and Jane Williams

External assessor

Tony Lawrence (University of Warwick)

Critical readers

Alison Brand and Shirley John

Curriculum manager

Gloria Baldi

Media project manager

Stephen Clift

Learning media developer

Lucinda Simpson

Graphics media developer

Jon Owen

Interactive media developer

Callum Lester

Media developer, Sound & Vision

Michael Francis

Licensing & acquisitions assistant

Carol Houghton

With the assistance of

Andy Allum, Val Aspland, Joe Buchanunn, Jim Campbell, Matt Compton, Sian Contell, Martin Keeling, Barbara Langley-Poole, Tara Marshall, Sandy Nicholson, Angela Noufaily, Daphne Turner, Andrew Whitehead and Kaye Williams

M140 is based on a former Open University module, MDST242 *Statistics in Society*. Much of MDST242 is used in M140, so the module team that produced MDST242 (acknowledged elsewhere) have also made substantive input to M140.

Solutions to activities

Solution to Activity 1

(a) There are 10 data values. The fifth longest time between elections is 49 months and the sixth longest is also 49 months. As

$$\frac{49+49}{2} = 49,$$

the median is 49 months.

(b) Using the formula,

$$\overline{x} = \frac{44+7+55+49+48+58+61+49+47+60}{10}$$
$$= \frac{478}{10} = 47.8.$$

So the mean is 47.8 months.

(c) One period between general elections was only 7 months, much smaller than the other values. This small value pulls down the mean but has little effect on the median. For this reason, the mean is smaller than the median. However, the mean and median are quite close and both seem reasonably representative of the data. (Though you might argue that 7 months is an outlier, and so the median is preferable to the mean because it is not affected by the odd outlier.)

Solution to Activity 2

(a) There are 15 data values, so

$$\frac{n+1}{4} = \frac{16}{4} = 4 \quad \text{and} \quad \frac{3(n+1)}{4} = \frac{48}{4} = 12.$$

Hence the lower quartile is the 4th item and the upper quartile is the 12th item (4th from the top). Therefore $Q_1 = 57$ and $Q_3 = 65$. The interquartile range is the distance between them: $65 - 57 = 8$.

(b) Now there are 20 data values, so

$$\frac{n+1}{4} = \frac{21}{4} = 5\tfrac{1}{4} \quad \text{and} \quad \frac{3(n+1)}{4} = \frac{63}{4} = 15\tfrac{3}{4}.$$

Thus the lower quartile is one-quarter of the way from 17 to 19 and the upper quartile is three-quarters of the way from 23 to 24. Therefore $Q_1 = 17.5$, $Q_3 = 23.75$ and the interquartile range is $23.75 - 17.5 = 6.25 \simeq 6.3$.

Solution to Activity 3

The values of $\sum x^2$ and $\sum x$ must be determined. Don't write down each number you square; just use your calculator memory:

$$\sum x^2 = 20.5^2 + 21.6^2 + \cdots + 23.6^2 = 4112.92,$$

and

$$\sum x = 20.5 + 21.6 + \cdots + 23.6 = 192.0.$$

The sample size is $n = 9$ and so the variance equals

$$s^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) = \frac{1}{8}\left(4112.92 - \frac{192.0^2}{9}\right)$$
$$= \frac{1}{8}(4112.92 - 4096.0)$$
$$= 2.115.$$

The standard deviation equals

$$\sqrt{\text{variance}} = \sqrt{2.115} \simeq 1.45.$$

Solution to Activity 4

(a) Reading from the figure, 4.5 kg is the 2nd percentile at 10 weeks old, so 2% of boys weigh less than 4.5 kg at 10 weeks.

(b) 10 kg is the 75th percentile at 46 weeks old, so 25% of boys weigh more than 10 kg at 46 weeks.

(c) 6 kg is the 2nd percentile at 22 weeks old, and 7.5 kg is the 50th percentile. Hence the proportion of 22-week-old boys weighing between 6 kg and 7.5 kg is $50\% - 2\% = 48\%$.

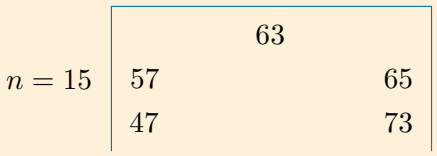
Solution to Activity 5

(a) From the solution to Activity 2, $Q_1 = 57$ and $Q_3 = 65$.

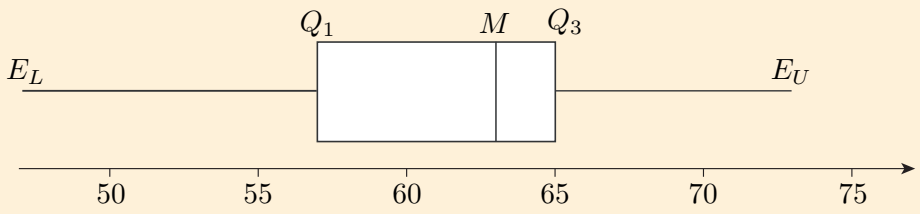
There are 15 data values, so the middle value is the 8th. Therefore, $n = 15$ and $M = 63$.

The lowest and highest values are $E_L = 47$ and $E_U = 73$.

Hence the following is the five-figure summary of the data:



(b) This is the boxplot:



(c) The median (M) is much nearer the upper quartile (Q_3) than the lower quartile (Q_1) and the left-hand whisker is a little longer than the right-hand whisker. Hence the boxplot shows the data are left-skew and, from the position of M, the skewness is fairly marked.

Solution to Activity 6

(a) There is one particularly high value, 53.46 seconds, which is listed separately. The stretched stemplot looks like this:

44	3	3
44	5	5 6 7 8 9
45	0	0 1 3 4
45	6	8 8
46	1	1 2
46	6	
47		
47	5	
48	1	
48	9	
HI	534	
$n = 24$ 44	3 represents 44.3 seconds	

(b) The stemplot shows that the data are clearly right-skew. There are a lot of values in the range 44–46.5 seconds, then a few values stretching out to 49 seconds, and then the value of 53.4 seconds.

Solution to Activity 7

For 'Food and catering',

$$rw = 1.013 \times 163 = 165.119.$$

Similar calculations for other groups yield the last column of the table.

Group	Price ratio for August 2013 relative to January 2013 (r)	2013 weights (w)	Price ratio $\times$ weight (rw)
Food and catering	1.013	163	165.119
Alcohol and tobacco	1.021	91	92.911
Housing and household expenditure	1.017	419	426.123
Personal expenditure	1.055	83	87.565
Travel and leisure	1.022	244	249.368

Then

$$\sum rw = 165.119 + 92.911 + \cdots + 249.368 = 1021.086$$

and

$$\sum w = 163 + 91 + \cdots 244 = 1000.$$

Thus the all-item price ratio is

$$\frac{\sum rw}{\sum w} = \frac{1021.086}{1000} \simeq 1.021.$$

Solution to Activity 8

(a) The numbers in each set are:

A	B	C	D	E	F
8	15	16	9	7	13

Dividing these by the population size of 68, the following gives the proportion of the population in each set:

A	B	C	D	E	F
0.118	0.221	0.235	0.132	0.103	0.191

(b) The labels selected are:

$$68,\ 42,\ 66,\ 14,\ 60,\ 63,\ 24,\ 06,\ 21,\ 52,\ 37,\ 19,\ 33,\ 55,\ 03,\ 11,\ 18.$$

The sample is shown in the following table:

Label	Initials	Set	Label	Initials	Set	Label	Initials	Set
03	A.P.D.	A	21	G.B.Y.	B	55	M.J.P.	E
06	J.L.	A	24	R.D.M.	C	60	M.A.T.	F
11	C.T.L.	B	33	D.J.S.	C	63	R.J.C.	F
14	J.S.R.	B	37	C.A.M.	C	66	E.D.	F
18	P.J.G.	B	42	A.L.	D	68	M.B.	F
19	W.W.S.	B	52	G.C.T.	E			

The numbers selected from each set are:

A	B	C	D	E	F
2	5	3	1	2	4

Dividing these by the sample size of 17, the following gives the proportion of the sample in each set:

A	B	C	D	E	F
0.118	0.294	0.176	0.059	0.118	0.235

Comparing these proportions with the proportions in solution (a), B and F are over-represented (and E is marginally over-represented), C and D are under-represented, while A is spot on.

(c) As we want to sample a quarter of the population, we will start at 1, 2, 3 or 4. The selected digit from row **10** is 3, so we start at label 03.

Every fourth label is in the sample:

$$03,\ 07,\ 11,\ 15,\ 19,\ 23,\ 27,\ 31,\ 35,\ 39,\ 43,\ 47,\ 51,\ 55,\ 59,\ 63,\ 67.$$

The sample is shown in the following table:

Label	Initials	Set	Label	Initials	Set	Label	Initials	Set
03	A.P.D.	A	27	T.P.H.	C	51	J.P.R.	E
07	J.D.H.	A	31	T.R.F.	C	55	M.J.P.	E
11	C.T.L.	B	35	D.L.	C	59	D.B.M.	F
15	S.L.	B	39	W.O.J.	C	63	R.J.C.	F
19	W.W.S.	B	43	L.R.P.	D	67	Y.S.H.	F
23	M.S.	B	47	Z.G.	D			

The numbers selected from each set are:

A	B	C	D	E	F
2	4	4	2	2	3

Dividing these by the sample size of 17, the following gives the proportion of the sample in each set:

A	B	C	D	E	F
0.118	0.235	0.235	0.118	0.118	0.176

The differences between these proportions and those in (a) are small. In fact, as each sample size must be a whole number they could not be closer. So the sample is a good representation of the population, as far as we can tell.

(d) The population is listed by set – first set A, then set B, ... Systematic sampling is picking every fourth person, so it is certain to take approximately a quarter of the people in each set. Hence it could be anticipated that the systematic sample would represent the population better than the simple random sample.

Solution to Activity 9

(a) There is no single correct answer to this question. One approach is to group parishes together so as to form a moderate number of strata – say about 10 strata. This is because there are too many parishes to treat each as an individual stratum. An alternative would be to treat the parishes as clusters and randomly sample 10 of them, say. (But forming clusters would not utilise knowledge about parishes regarding their geographic location, proportion of council houses, or age of houses in different areas.) Each stratum should consist of adjacent parishes that have a similar housing stock.

Houses within a stratum should be grouped by their council tax band and by whether or not they are privately owned. (Some council tax bands might be combined to reduce the number of strata.) Then streets within each of these strata should be treated as clusters and a random subset of them selected. It is essential to treat streets as clusters so as to reduce the travelling time of the interviewer.

A large number of houses should be selected (perhaps 50%) from each street, as only a small proportion will have any disabled people living in them. Most of the interviews will consequently be very fast, so walking between interviews should be minimised.

(b) The above procedure is likely to yield a sample that is better than a completely random sample at reflecting the characteristics of the unitary authority's housing stock. This should also hold for alternative sampling schemes that you may have proposed. Hence the procedure should reduce sampling variation compared with that of simple random sampling. The main benefit though, is that travelling time should be far less because only a modest number of streets (the clusters) will feature in the survey, so the survey should cost much less than with a simple random sample.

Solution to Activity 10

Statement G is true: in a double-blind trial, neither the patients nor the doctors know which patients have received the drug being tested and which the placebo, while an appropriate independent person does have this information. Consequently, statements A and D are false.

It has been shown that patients can genuinely respond to the process of being treated, even when the treatment contains no beneficial ingredient. Similarly, doctors can influence a patient's response in ways other than through the medication. Hence statement E is true, while B and C are false.

Patients must give informed consent before they are entered into a clinical trial. Thus they must always be told that they might be given just a placebo. Indeed, it would be unethical not to tell them that. Thus statement F is false.

Solution to Activity 11

(a) It is a group-comparative design because there are two groups, patients were allocated to the experimental group or control group at random and the design had no other features or additional complexity.

(b) The validity of the trial is questionable because any relevant gender differences would bias results. For example, if men tended to have more extreme arthritis symptoms than women, then the experimental treatment (taken by the women) would only seem poorer than the existing drug (taken by the men) if it was actually *much* worse than the existing drug.

(c) If the experiment is repeated, half the women should be allocated at random to the experimental group and half to the control group, using a simple random sample to choose which women were allocated to the experimental group. Similarly, half the men should be picked at random from the male patients and allocated to the experimental group and the other half of the men should be allocated to the control group. This meets the requirements of randomisation (so neither drug is deliberately favoured) and ensures the same gender ratio in the two groups. Essentially, this procedure treats women as one strata and men as a second strata – it is called *stratified randomisation.*

Solution to Activity 12

(a) There are 227 research fellows in the population of 794 academic statisticians. Hence, the probability that the selected person is a research fellow is

$$\frac{227}{794} \simeq 0.286.$$

(b) Among the 794 academic statisticians, there are 275 aged 30–39. Hence, the probability that the selected person is aged 30–39 is

$$\frac{275}{794} \simeq 0.346.$$

(c) Among the 794 academic statisticians, there are 82 research fellows aged 30–39. Hence, the probability that the selected person is a research fellow aged 30–39 is

$$\frac{82}{794} \simeq 0.103.$$

Solution to Activity 13

(a) The subpopulation of interest are the 211 people in Table 4 who are aged 40–49, of whom 70 are senior lecturers. Hence,

$$P(\text{senior lecturer}|\text{aged 40–49}) = \frac{70}{211} \simeq 0.332.$$

(b) Now the subpopulation of interest are the 156 people in Table 4 who are senior lecturers. Of these people, 70 are aged 40–49. Hence,

$$P(\text{aged 40–49}|\text{senior lecturer}) = \frac{70}{156} \simeq 0.449.$$

Solution to Activity 14

(a) There are 66 professors aged 50–59. As there are 794 academic statisticians,

$$P(\text{professor aged 50–59}) = \frac{66}{794} \simeq 0.083.$$

(b) There are 172 professors in total. Hence,

$$P(\text{professor}) = \frac{172}{794} \quad (\simeq 0.217).$$

(c) The subpopulation of interest are the 172 people in Table 4 who are professors, of whom 66 are aged 50–59. Hence,

$$P(\text{aged 50–59}|\text{professor}) = \frac{66}{172} \quad (\simeq 0.384).$$

(d) From (a),

$$P(A \text{ and } B) = P(\text{professor aged 50–59}) \simeq 0.083.$$

Also, from (b) and (c),

$$P(A) = P(\text{professor}) = \frac{172}{794}$$

and

$$P(B|A) = P(\text{aged 50–59}|\text{professor}) = \frac{66}{172}.$$

So

$$\begin{aligned} P(A) \times P(B|A) &= \frac{172}{794} \times \frac{66}{172} \\ &= \frac{66}{794} \\ &\simeq 0.083 = P(A \text{ and } B). \end{aligned}$$

(In decimals, $P(A) \times P(B) \simeq 0.217 \times 0.384 \simeq 0.083$.)

Solution to Activity 15

(a) From Table 4, there are 172 professors and $7 + 6 + 8 = 21$ so there are 21 academic statisticians aged at least 60 who are not professors. Hence, adding 172 to 21 gives a total of 193 UK academic statisticians who will be invited. If we simply added the number of professors (172) to the number of academic statisticians aged 60 or over (68), then we would double-count the 47 people who are both professors *and* aged 60 or over. Another way of obtaining the figure of 193 is from $172 + 68 - 47 = 193$.

(b) There are 47 UK academic statisticians who are both professors *and* aged 60 or over. So 47 people would be invited.

(c) From (a),

$$P(A \text{ or } B) = \frac{193}{794} \simeq 0.243$$

and from (b),

$$P(A \text{ and } B) = \frac{47}{794} \simeq 0.059.$$

(d) $$P(A) = \frac{172}{794} \simeq 0.217 \quad \text{and} \quad P(B) = \frac{68}{794} \simeq 0.086.$$

Hence,

$$\begin{aligned} P(A) + P(B) - P(A \text{ and } B) &\simeq 0.217 + 0.086 - 0.059 \\ &= 0.244 \\ &\simeq P(A \text{ or } B). \end{aligned}$$

Alternatively, to avoid having to deal with rounding you could put

$$\begin{aligned} P(A) + P(B) - P(A \text{ and } B) &= \frac{172}{794} + \frac{68}{794} - \frac{47}{794} \\ &= \frac{172 + 68 - 47}{794} \\ &= \frac{193}{794} \\ &= P(A \text{ or } B). \end{aligned}$$

Solution to Activity 16

(a) The number of ways of choosing three pairs from six (order does not matter) is

$${}^6C_3 = \frac{6 \times 5 \times 4}{3 \times 2 \times 1} = 20.$$

(b) The number of ways of choosing four pairs from nine is

$${}^9C_4 = \frac{9 \times 8 \times 7 \times 6}{4 \times 3 \times 2 \times 1} = 126.$$

Solution to Activity 17

(a)

$$\begin{aligned} P(FSSSFS) &= 0.4 \times 0.6 \times 0.6 \times 0.6 \times 0.4 \times 0.6 \\ &= 0.6^4 \times 0.4^2 \\ &= 0.1296 \times 0.16 \\ &= 0.020\,736. \end{aligned}$$

(b) The number of sequences of 6 days that give 4 successes is

$${}^6C_4 = \frac{6 \times 5 \times 4 \times 3}{4 \times 3 \times 2 \times 1} = 15.$$

(c)

$$\begin{aligned} P(\text{4 successful days in 6 days}) &= {}^6C_4 \times 0.6^4 \times 0.4^2 \\ &= 15 \times 0.020\,736 \\ &= 0.311\,04 \simeq 0.311. \end{aligned}$$

(d)

$$\begin{aligned} P(\text{5 successful days in 8 days}) &= {}^8C_5 \times 0.6^5 \times 0.4^3 \\ &= \frac{8 \times 7 \times 6 \times 5 \times 4}{5 \times 4 \times 3 \times 2 \times 1} \times 0.6^5 \times 0.4^3 \\ &= 56 \times 0.077\,76 \times 0.064 \simeq 0.279. \end{aligned}$$

Solution to Activity 18

Define 'blood group O' as success, so that $p = P(S) = 0.44$ and $q = 1 - p = 0.56$. The number of trials (number of people) is $n = 7$ and we want $P(x = 2)$. Using the formula for the binomial distribution,

$$\begin{aligned} P(x = 2) &= {}^7C_2 \times 0.44^2 \times 0.56^{(7-2)} \\ &= {}^7C_2 \times 0.44^2 \times 0.56^5 \\ &\simeq \frac{7 \times 6}{2 \times 1} \times 0.1936 \times 0.055\,073 \\ &= 21 \times 0.010\,662 \\ &\simeq 0.224. \end{aligned}$$

Solution to Activity 19

One reason for not using $\sum(\text{Residual})^2$ as the test statistic is that a residual's evidence against H_0 depends not only on the size of the Residual, but also on the size of the Expected value. To illustrate, suppose that a cell is expected to contain 3400 items but in fact contains 3420. The difference is relatively small – readily explained as random variation. On the other hand, if a cell is expected to contain 10 items but in fact contains 30, then the difference is large. Thus, although both cases give a $(\text{Residual})^2$ of $20^2 = 400$, only the latter provides strong evidence that the Expected value is wrong.

Another reason (which you will not be aware of) is that we do not know the probability distribution of $\sum(\text{Residual})^2$. As noted earlier (point 3 in the description of a hypothesis test), the probability distribution of a suitable test statistic must be fully known if H_0 is true.

Solution to Activity 20

(a) When $\mu = 0.38$ and $\sigma = 0.03$, the formula

$$z = \frac{x - \mu}{\sigma} \text{ gives } z = \frac{x - 0.38}{0.03}.$$

When $x = 0.40$,

$$z = \frac{0.40 - 0.38}{0.03} = \frac{0.02}{0.03} \simeq 0.667.$$

So a shell thickness of 0.40 mm is 0.667 standard deviations above the mean thickness of 0.38 mm.

(b) When $x = 0.30$,

$$z = \frac{0.30 - 0.38}{0.03} = \frac{-0.08}{0.03} \simeq -2.667.$$

So a shell thickness of 0.30 mm is 2.667 standard deviations below the mean thickness of 0.38 mm.

Solution to Activity 21

(a) The sampling distribution of the mean mark for a sample of size 5 has mean 66 and standard deviation $\sigma/\sqrt{n} = 22/\sqrt{5} \simeq 9.83$.

The sampling distribution of the mean mark for a sample of size 15 also has mean 66, but its standard deviation is $22/\sqrt{15} \simeq 5.68$.

(b) The sample sizes of 5 and 15 are not large, so the distributions of the sample mean may differ a little from a normal distribution – more so with the smaller sample size of 5.

Solution to Activity 22

(a) The variance is known so, from the flow chart in Figure 9, use the z-test.

(b) The variance is unknown and the sample size is greater than 25 so, from Figure 9, use the z-test.

(c) The variance is unknown and the sample size is less than 25 so, from Figure 9, use the t-test.

(d) The variance is known so, from Figure 9, use the z-test.

Solution to Activity 23

(a) The data are not matched pairs, the population variances are unknown, and both sample sizes are above 25 so, from the flow chart in Figure 10, use the two-sample z-test.

(b) The data are not matched pairs, the population variances are unknown, one sample size is less than 25, but we can assume the population variances are equal ($9.6/8.4 \simeq 1.14 < 3$) so, from Figure 10, use the two-sample t-test with a pooled sample variance.

(c) The data are matched pairs so, from Figure 10, use the matched-pairs t-test.

(d) The data are not matched pairs, the population variances are unknown, one sample size is less than 25 (in fact, both are less than 25), and we cannot assume the population variances are equal ($12.1/3.5 \simeq 3.46 > 3$) so, from Figure 10, use the two-sample t-test with unequal variances.

(e) The data are not matched pairs and the population variances are known so, from the flow chart in Figure 10, use the two-sample z-test.

Solution to Activity 24

Let μ_A and μ_B denote the population mean reductions in cholesterol level on the oat and bean diets (diets A and B), respectively. Then

$$n_A = 29, \quad \overline{x}_A = 58.3, \quad s_A = 19.2,$$
$$n_B = 28, \quad \overline{x}_B = 46.4, \quad s_B = 16.5.$$

The data are not paired, the population variances are not known, and both sample sizes are above 25. From Figure 10, a two-sample z-test is appropriate.

The hypotheses are:

$$H_0\colon \mu_A = \mu_B \quad \text{and} \quad H_1\colon \mu_A \neq \mu_B.$$

We first calculate the value of ESE, the estimated standard error of $\overline{x}_A - \overline{x}_B$:

$$\text{ESE} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{19.2^2}{29} + \frac{16.5^2}{28}} \simeq 4.7366.$$

Hence the value of the test statistic is

$$z = \frac{\overline{x}_A - \overline{x}_B}{\text{ESE}} \simeq \frac{58.3 - 46.4}{4.7366} \simeq 2.512.$$

The critical values are ± 1.96 (at 5%) and ± 2.58 (at 1%). Since $1.96 < 2.512 < 2.58$, we can reject H_0 at the 5% significance level but not at the 1% significance level. There is moderate evidence that the two diets differ in the average reduction in serum cholesterol level that they yield. The reduction on the oat diet (diet A) appears to be greater.

The only assumptions needed for this hypothesis test are that observations are all random and independent. The sample sizes are quite large, so the central limit theorem implies that the distribution of $\overline{x}_A$ will be approximately normal, as will that of $\overline{x}_B$.

Solution to Activity 25

(a) The point estimate is $\overline{x}_A - \overline{x}_B = 58.3 - 46.4 = 11.9$. From the solution to Activity 24, ESE $\simeq 4.7366$. A two-sample z-test was used in Activity 24 so z-values will be the critical values.

For a 95% confidence interval, the z-value is 1.96. Hence the lower limit of the 95% confidence interval is

$$11.9 - 1.96 \times 4.7366 \simeq 2.62$$

and the upper limit is

$$11.9 + 1.96 \times 4.7366 \simeq 21.18.$$

Thus the 95% confidence interval for $\mu_A - \mu_B$ is $(2.62\,\text{mg/dl},\ 21.18\,\text{mg/dl})$.

(b) For a 99% confidence interval, the z-value is 2.58. Hence the lower limit of the 99% confidence interval is

$$11.9 - 2.58 \times 4.7366 \simeq -0.32$$

and the upper limit is

$$11.9 + 2.58 \times 4.7366 \simeq 24.12.$$

Thus the 99% confidence interval for $\mu_A - \mu_B$ is $(-0.32\,\text{mg/dl},\ 24.12\,\text{mg/dl})$.

(c) A 95% confidence interval is always shorter than the corresponding 99% interval, as in this example.

(d) The 99% confidence interval contains 0 while the 95% confidence interval does not contain 0. Hence the hypothesis H_0: $\mu_A - \mu_B = 0$ would not be rejected at the 1% significance level but it would be rejected at the 5% significance level. This was the result found in Activity 24.

Solution to Activity 26

In (a), the points all lie very close to a straight line sloping downwards, so there is a strong negative linear relationship between the two variables.

In (b), the points follow a general upward trend, but a smooth line could not be drawn through the points that lay close to many of them. Thus there is a weak positive relationship between the two variables – it may be linear, but that is not clear.

In (c), the scatterplot suggests no relationship between the two variables.

In (d), the points follow a clear downward trend, but most of them would be some way from a straight line drawn through them. Thus there is a negative relationship between the two variables that looks linear, but is not very strong.

In (e), the points all lie very close to a straight line sloping upwards, so there is a strong positive linear relationship between the two variables.

In (f), the points all lie very close to a line that increases from left to right, so there is a strong positive relationship between the two variables. However, the points clearly follow a curve (not a straight line) so the relationship is non-linear.

Solution to Activity 27

From the scatterplots:

- A strong positive linear relationship is shown in (e).
- A strong positive non-linear relationship is shown in (f).
- A weak positive relationship is shown in (b).
- No relationship is shown in (c).
- A weak negative relationship is shown in (d).
- A strong negative linear relationship is shown in (a).

From the figures it is not transparent whether (b) will give a higher correlation coefficient than (f), or vice versa. In fact, (f) has the higher correlation coefficient. The correct ordering is as follows. Correlation coefficients are also given. (You do not have the information to calculate the correlations.)

(e): 0.99, (f): 0.88, (b): 0.77, (c): 0.13, (d): -0.82, (a): -0.99.

Solutions to exercises

Solution to Exercise 1

(a) We start in row **36** of the random number table.

For the first stratum, we want six numbers between 01 and 23:

$$18,\ 17,\ 19,\ 08,\ 01,\ 16.$$

Rearranged in order, the labels are:

$$01,\ 08,\ 16,\ 17,\ 18,\ 19.$$

For the second stratum, we want six numbers between 24 and 48. Starting in the random number table at the point where we left off, we get:

$$31,\ 42,\ 25,\ 37,\ 35,\ 44.$$

Rearranged in order, the labels are:

$$25,\ 31,\ 35,\ 37,\ 42,\ 44.$$

For the third stratum, we want five numbers between 49 and 68. Continuing from where we left off in the random number table, we get:

$$61,\ 55,\ 64,\ 68,\ 56.$$

Rearranged in order, the labels are:

$$55,\ 56,\ 61,\ 64,\ 68.$$

(b) The first two single digits between 1 and 6 in row **95** are 4 and 6, which correspond to set D and set F.

Set D is of size 9, so we select $9/3 = 3$ people from set D. Set F is of size 13 and $13/3$ rounds up to 5, so we select five people from set F.

Starting in row **72**, for set D we want three random number pairs between 40 and 48:

$$42,\ 48,\ 46.$$

Continuing on, for set F we want five numbers between 56 and 68:

$$61,\ 59,\ 68,\ 65,\ 60.$$

Hence, the following are the people in the subsamples (rearranged in the order of their labels):

- from set D: A.L., A.T. and Y.H.
- from set F: D.B.M., M.A.T., A.N.D., G.K.S. and M.B.

Solution to Exercise 2

(a) Age-bands should be formed (perhaps under 25, 25–34, 35–44, 45–54, 55 or over), and job grades should be grouped together (perhaps forming five or six groups) so that similar grades are in the same group.

Employees should be listed so that those who share the same age-band and grade-group are listed consecutively. Each age-band/grade-group combination is a stratum.

A sample of 500 from 6000 means that a 12th of the workforce is to be included in the sample. Generate a random number between 1 and 12 (either using a computer or random number tables). Pick out every 12th

person from the list, starting with the person given by the random number. This gives a systematic sample and the proportion of people chosen from each stratum will be approximately the same for each stratum.

(b) Non-sampling errors could arise from refusal to answer, employee absence, out-of-date data, incorrect (or incorrectly transcribed) data, etc.

Solution to Exercise 3

(a) For a crossover design, a random half of the women should take the new treatment for four weeks. They should then switch to the old treatment. The difference of each patient's symptoms under the two treatments should be recorded. The other half of the women should start on the old treatment for four weeks and then switch to the new treatment. The difference in response between new and old treatments is again the quantity of interest. The same should be done with the men. This design means that order of treatment should not bias results, and gender-bias is also controlled. A reasonable time on each treatment should elapse before symptoms are measured, so that they reflect the current treatment.

(b) For a matched-pairs design, pairs of patients are required. The patients within a pair should be similar with respect to gender, age and severity of illness. Then one person in the pair would be picked at random and allocated to the experimental treatment and the other person in the pair would receive the control treatment. The difference in their symptoms would be the data analysed.

(c) Finding matched pairs is hard, while a crossover trial is straightforward to run because arthritis is a chronic condition. Hence the crossover design is more suitable than a matched-pairs design. When they are suitable, crossover designs are better than group-comparative designs because they remove much of the variation between individuals as each individual is compared with himself or herself. Hence a group-comparative design would not be better.

Solution to Exercise 4

(a) Define events A and B as follows.

A: a randomly picked academic statistician is aged 40–49.

B: a randomly picked academic statistician is a lecturer.

Then

$$P(A) = \frac{211}{794} \simeq 0.266 \quad \text{and} \quad P(A|B) = \frac{59}{239} \simeq 0.247.$$

As $P(A)$ does not equal $P(A|B)$, events A and B are not independent.

(b) As

$$P(B) = \frac{239}{794} \simeq 0.301 \quad \text{and} \quad P(A|B) \simeq 0.247,$$

then

$$P(A \text{ and } B) \simeq 0.301 \times 0.247 \simeq 0.074.$$

(c) Using the answers from (a) and (b),

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) - P(A \text{ and } B) \\ &\simeq 0.266 + 0.301 - 0.074 = 0.493. \end{aligned}$$

Solution to Exercise 5

(a) The number of successes follows a binomial distribution. As $p = 0.2$, $q = 1 - p = 0.8$.

Here $n = 4$, so

$$P(x = 1) = {}^4C_1 \times 0.2^1 \times 0.8^{(4-1)}$$
$$= \frac{4}{1} \times 0.2 \times 0.8^3 = 0.4096 \simeq 0.410.$$

(b) Now $n = 5$, so

$$P(x = 2) = {}^5C_2 \times 0.2^2 \times 0.8^{(5-2)}$$
$$= \frac{5 \times 4}{2 \times 1} \times 0.2^2 \times 0.8^3 = 0.2048 \simeq 0.205.$$

(c) Now $n = 7$, so

$$P(x = 0) = {}^7C_0 \times 0.2^0 \times 0.8^{(7-0)}$$
$$= 0.8^7 \simeq 0.210,$$

as ${}^7C_0 = 1$ and $0.2^0 = 1$.

Solution to Exercise 6

(a) It must be assumed that whether one person feels better is independent of whether other people feel better. Then probabilities come from a binomial distribution with $n = 6$, $p = 0.6$ and $q = 1 - p = 0.4$.

$$P(x = 2) = {}^6C_2 \times 0.6^2 \times 0.4^4$$
$$= \frac{6 \times 5}{2 \times 1} \times 0.36 \times 0.0256$$
$$= 0.138\,24.$$

(b) Using the same assumptions as in (a),

$$P(x = 1) = {}^6C_1 \times 0.6^1 \times 0.4^5$$
$$= \frac{6}{1} \times 0.6 \times 0.010\,24$$
$$\simeq 0.036\,86$$

and

$$P(x = 0) = {}^6C_0 \times 0.6^0 \times 0.4^6$$
$$= 0.4^6 \simeq 0.004\,10.$$

Thus P(2 or fewer people feeling better) is

$$P(x = 2) + P(x = 1) + P(x = 0) \simeq 0.138\,24 + 0.036\,86 + 0.004\,10$$
$$\simeq 0.179.$$

Solution to Exercise 7

(a) The hypotheses are as follows:

H_0: locality and concern about air pollution are independent.

H_1: locality and concern about air pollution are not independent.

(b) Copy marginal totals from the Observed table to the Expected table.

Then the Expected value for the first cell is:

$$\frac{\text{Row total} \times \text{Column total}}{\text{Overall total}} = \frac{100 \times 83}{400}$$
$$= 20.75.$$

The other values are obtained in the same manner, leading to the following Expected table:

Locality	Yes	No	Total
A	20.75	79.25	100
B	20.75	79.25	100
C	20.75	79.25	100
D	20.75	79.25	100
Total	83	317	400

As a check on your calculations, the Expected values within the table must add to the marginal totals (apart from unimportant rounding errors).

(c) The Residual table is found by subtracting the terms of the Expected table from those of the Observed table. For the first cell, we have

$$25 - 20.75 = 4.25.$$

The Residual table is

Locality	Yes	No
A	4.25	-4.25
B	-4.75	4.75
C	-8.75	8.75
D	9.25	-9.25

The χ^2 contribution of the first cell is

$$\frac{(\text{Residual})^2}{\text{Expected}} = \frac{4.25^2}{20.75} \simeq 0.8705.$$

The complete table of χ^2 contributions is:

Locality	Yes	No
A	0.8705	0.2279
B	1.0873	0.2847
C	3.6898	0.9661
D	4.1235	1.0797

(d) The χ^2 test statistic is the sum of the eight χ^2 contributions:

$$\begin{aligned}\chi^2 &= 0.8705 + 0.2279 + 1.0873 + 0.2847 + 3.6898 \\ &\quad + 0.9661 + 4.1235 + 1.0797 \\ &\simeq 12.330.\end{aligned}$$

The Observed table is a 4×2 table, so its degrees of freedom are $(4-1) \times (2-1) = 3$. Hence $\text{CV}5 = 7.815$ and $\text{CV}1 = 11.345$.

(e) Since $12.330 > 11.345$, we reject the null hypothesis at the 1% significance level. Thus there is strong evidence that concern in a household about air pollution varies with locality.

Solution to Exercise 8

(a) The variance is known so, from the flow chart in Figure 9 (Subsection 5.2), use the z-test.

(b) The variance is unknown and the sample size is less than 25 so, from Figure 9, use the t-test.

(c) The variance is unknown and the sample size is greater than 25 so, from Figure 9, use the z-test.

(d) The variance is known so, from Figure 9, use the z-test.

Solution to Exercise 9

(a) The data are not matched pairs, the population variances are unknown, one sample size is less than 25 (in fact, both are less than 25), but we can assume the population variances are equal ($28.2/23.8 \simeq 1.18 < 3$) so, from the flow chart in Figure 10, use the two-sample t-test with a pooled sample variance.

(b) The data are matched pairs so, from Figure 10, use the matched-pairs t-test.

(c) The data are not matched pairs, the population variances are unknown, one sample size is less than 25, and we cannot assume the population variances are equal ($9.1/1.7 \simeq 5.35 > 3$) so, from Figure 10, use the two-sample t-test with unequal variances.

(d) The data are not matched pairs, the population variances are unknown, and both sample sizes are above 25 so, from Figure 10, use the two-sample z-test.

(e) The data are not matched-pairs, the population variances are unknown, one sample size is less than 25, and we can assume the population variances are equal ($17.4/10.6 \simeq 1.64 < 3$) so, from Figure 10, use the two-sample t-test with a pooled sample variance.

Solution to Exercise 10

(a) The data give:

$$\begin{aligned}\sum(x-\overline{x})^2 &= \sum x^2 - \frac{(\sum x)^2}{n}\\ &\simeq 19\,378 - \frac{(620)^2}{30} \simeq 6564.7.\\ \sum(y-\overline{y})^2 &= \sum y^2 - \frac{(\sum y)^2}{n}\\ &= 344\,742 - \frac{(2776)^2}{30} \simeq 87\,869.5.\\ \sum(x-\overline{x})(y-\overline{y}) &= \sum xy - \frac{(\sum x)(\sum y)}{n}\\ &= 79\,136 - \frac{620\times 2776}{30} \simeq 21\,765.3.\end{aligned}$$

Hence,

$$\begin{aligned}\text{correlation} &= \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \times \sum(y-\overline{y})^2}}\\ &\simeq \frac{21\,765.3}{\sqrt{6564.7\times 87\,869.5}} \simeq 0.91.\end{aligned}$$

(b) The correlation coefficient is quite large, so a straight line should be useful for capturing the relationship between time at the checkout and number of purchases. However, the relationship might be non-linear, in which case there would be better ways of capturing the relationship.

The data on each individual customer is needed, so that a scatterplot of y against x could be drawn. That would enable the relationship between the variables to be seen much more clearly – the question of whether a straight-line relationship is appropriate could then be answered with greater confidence.

Solution to Exercise 11

From Exercise 10,

$$\sum x = 620, \quad \sum y = 2776, \quad \text{and} \quad n = 30.$$

So

$$\overline{x} = \frac{\sum x}{n} = \frac{620}{30} \simeq 20.667$$

and

$$\overline{y} = \frac{\sum y}{n} = \frac{2776}{30} \simeq 92.533.$$

Also, from the solution to Exercise 10,

$$\sum (x - \overline{x})^2 \simeq 6564.7 \quad \text{and} \quad \sum (x - \overline{x})(y - \overline{y}) \simeq 21\,765.3.$$

Hence, the slope is

$$b = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sum (x - \overline{x})^2} \simeq \frac{21\,765.3}{6564.7} \simeq 3.3155,$$

and the intercept is

$$\begin{aligned} a = \overline{y} - b\overline{x} &= 92.533 - 3.3155 \times 20.667 \\ &= 92.533 - 68.521 \simeq 24.012. \end{aligned}$$

These give the regression line:

$$y = 24.0 + 3.32x.$$

The intercept of 24.0 implies the basic part of serving a customer (taking money and giving change, etc.) takes 24.0 seconds, on average. The slope of 3.32 means that each item takes 3.32 seconds to process, on average.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Figure 2: Reproduced with permission of the Royal College of Paediatrics and child health.

Introduction, photo © Ratina Thongteeka/Dreamstime.com

Subsection 1.1 photo: Courtesy of Waveney District Council

Subsection 1.2 cartoon: http://xkcd.com/833/. This file is licensed under the Creative Commons Attribution-Non-commercial Licence http://creativecommons.org/licenses/by-nc/3.0/

Subsection 1.3 photo: Courtesy Flotrack.org

Subsection 2.1 cartoon: www.causeweb.org

Subsection 2.1 photo (cluster sampling): Howard Stanbury / http://www.flickr.com/photos/27195496@N00/2725957867. This file is licensed under the Creative Commons Attribution-Non-commercial-ShareAlike Licence http://creativecommons.org/licenses/by-nc-sa/3.0/

Subsection 2.2 photo, taken from: http://rogueestate.com/2009/11/09/the-curried-bacon-experiment/

Subsection 2.2 cartoon, taken from: http://stats.stackexchange.com/questions/423/what-is-your-favorite-data-analysis-cartoon

Subsection 3.1 photo, used with permission of the Statistical Laboratory, University of Cambridge

Subsection 3.1 cartoon, www.causeweb.org

Section 4 photo (left-hand image) © Konradbak/Dreamstime.com

Section 4 photo (right-hand image): With kind permission of https://multimedia.actiononhearingloss.org.uk/

Subsection 5.1 cartoon, www.causeweb.org

Subsection 5.1 photo, United States Department of Agriculture / http://en.wikipedia.org/wiki/File:Poultry_Classes_Blog_photo_-_Flickr_-_USDAgov.jpg. This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Subsection 5.2 photo (Brinell hardness testing machine), taken from: www.mltest.com/images/stories/shb-3000c.jpg

Subsection 5.2 photo (cookies), taken from: www.fitsugar.com/Healthy-Oatmeal-Cookie-Recipe-Using-Beans-22665129

Subsection 5.2 figure (confidence intervals guide), taken from: http://excelmasterseries.com

Subsection 6.1 photo (Hans Rosling): With permission of Hans Rosling

Subsection 6.1 cartoon, www.cartoonstock.com

Subsection 6.2 photo (Sitka spruce): Wsiegmund. This file is licensed under the Creative Commons Attribution-Share Alike Licence http://creativecommons.org/licenses/by-sa/3.0/

Exercises on Section 6, photo: Veleta. This file is licensed under the Creative Commons Attribution-Share Alike Licence http://creativecommons.org/licenses/by-sa/3.0/

Learning outcomes, photo: Ethan Prater. This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked the publishers will be pleased to make the necessary arrangements at the first opportunity.

Index